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

***Российская Федерация***

**148 Интеллектуальные информационные системы и технологии. Лабораторная 6. Исследование моделируемой предметной области 2020-10-20**

**Заголовок:**
Лаб. №6: Исследование предметной области методами АСК-анализа (портреты, SWOT, кластеры, нейросети)

**Резюме:**

**1. Введение и контекст:**
Лекция и лабораторная работа №6 для группы ИТ-18-22 по дисциплине "Интеллектуальные информационные системы и технологии" состоялась 20 октября 2020 года. Занятие проводили профессора Луценко Е.В. и Аршинов Г.А. Тема занятия – исследование предметной области.

**2. Основная концепция: Исследование предметной области через моделирование:**
Исследование предметной области осуществляется путем анализа ее моделей. Система (предположительно, АСК-анализ, система "Эйдос") позволяет синтезировать модель на основе данных и затем исследовать эту модель различными методами.

**3. Ключевые методы анализа, продемонстрированные в системе:**

* **Информационные портреты классов:** Показывают наиболее и наименее характерные признаки (свойства, факторы) для объектов данного класса. Включают позитивный (характерные признаки) и негативный (нехарактерные признаки) портреты. Основаны на концепции конструктов (биполярных шкал), разработанной автором системы еще в 1987 году.
* **SWOT-анализ:** Представляет ту же информацию, что и информационные портреты, но в стандартной для экономистов форме (сильные/слабые стороны – внутренние факторы; возможности/угрозы – внешние факторы). Позволяет визуализировать факторы, способствующие или препятствующие достижению целевого состояния.
* **Когнитивные диаграммы и нейросетевая визуализация:** Визуализируют образ класса в виде нейрона, где свойства – рецепторы. Показывают активирующие (красные) и тормозящие (синие) связи. Интегральная 3D-карта отображает всю нейронную сеть модели, показывая связи между всеми признаками (рецепторами) и классами (нейронами), их силу (толщина линии) и знак (цвет). Позволяет оценить степень детерминированности классов.
* **Кластерно-конструктивный анализ:** Основан на матрице сходства классов (или признаков). Визуализируется через:
	+ Плоские когнитивные диаграммы сходства/различия.
	+ Агломеративные иерархические дендрограммы, показывающие процесс объединения классов/признаков в кластеры на основе их сходства. Система использует уникальный алгоритм кластеризации с пересчетом модели на каждом шаге, что считается более корректным, чем усреднение в классических методах.
* **Анализ ценности признаков/шкал:** Система рассчитывает информативность (ценность) каждого признака или шкалы для решения задачи идентификации/дифференциации классов. Позволяет выявить наиболее значимые и наименее полезные (или бесполезные) признаки/шкалы, что важно для оптимизации модели и понимания предметной области. Используется концепция опосредованных правдоподобных рассуждений для восстановления информации по коррелирующим признакам.

**4. Задание для студентов:**
Студентам дано комплексное задание:

* Выбрать тему приложения.
* Найти данные (рекомендуются репозитории UCI, Kaggle или собственные данные).
* Разработать приложение в системе "Эйдос".
* Зарегистрироваться в ResearchGate и РИНЦ (eLIBRARY), получить Spin-код, заключить договор с eLIBRARY.
* Разместить описание приложения в ResearchGate, РИНЦ и облаке "Эйдос".
* Оценка будет производиться по четким критериям, включая выполнение всех пунктов регистрации и размещения. Невыполнение задания или его частей приведет к снижению оценки.

**5. Ключевые моменты:**

* Подчеркивается важность корректной обработки данных в разных шкалах и единицах измерения, что решается в системе через расчет количества информации.
* Критикуются стандартные подходы кластерного анализа за некорректные допущения (например, усреднение матрицы сходства).
* Акцентируется внимание на необходимости активного участия студентов и своевременного выполнения задания.

**Детальная расшифровка текста:**

**1. Введение и контекст**

Здравствуйте, Георгий Александрович. Здравствуйте, ребята.

Сегодня у нас 20 октября 2020 года. Третья пара сейчас. 11:30-13:00 она продолжается. Эта пара. Лабораторная работа номер шесть с группой ИТ-18-22 по дисциплине "Интеллектуальные информационные системы и технологии".

Занятия ведут профессор Луценко Евгений Вениаминович и профессор Аршинов Георгий Александрович.

**2. Основная концепция: Исследование предметной области через моделирование**

Сегодня у нас по плану исследование предметной области. То есть мы рассматриваем вопрос об исследовании предметной области.

Что входит в это понятие? Исследование предметной области путём исследования модели этой области.

3. Демонстрация работы в системе "Эйдос"

3.1. Установка лабораторной работы и синтез модели

Сейчас я устанавливаю лабораторную работу и показываю в ней, как осуществляется исследование предметной области.

Что-то как-то вас маловато, ребята, нет? Или мне так кажется просто, что маловато.

Вы сейчас должны смотреть, как я устанавливаю работу. Я в режиме 1.3 выбираю приложение 3.03, добавить лабораторную работу третьего типа, третью работу 3.03. И сразу же после её добавления, вот сейчас этот процесс осуществляется. Сейчас у меня компьютер на 90% процессор загружен другой задачей. То есть он сейчас прямо пашет. Вот. И осуществляем синтез модели.

Это необходимо для того, чтобы исследовать модель. Исследовать объект моделирования путём исследования модели, нужно сначала создать модели. Но мы видим, что это быстро происходит. Модель очень маленькая. Такая демонстрационная модель.

3.2. Компоненты исследования предметной области

И начинаем заниматься исследованием моделируемой предметной области. Что сюда входит, ребята? Сюда входят следующие вопросы: типологический анализ классов. Сюда входят информационные портреты классов, кластерный конструктивный анализ классов, когнитивные диаграммы классов, а также входит сюда SWOT-анализ.

3.3. Информационные портреты классов

Смотрим на информационные портреты классов. Что они собой представляют? Берём, выбираем здесь в режиме 4.2.1, выбираем какой-то класс, выбираем нужную модель. В данном случае мы выбираем модель INF3, как наиболее достоверную. И мы справа видим, какие значения свойств или факторов наиболее характерны для объектов этого класса. И степень характерности видим. А потом идёт в порядке уменьшения, потом видим менее характерные, потом вообще видим нехарактерные. И дальше идёт наиболее нехарактерные. И мы видим, что все эти значения свойств, они образуют собой конструкт.

Я вот сейчас нажимаю кнопочку "вписать в окно", и мы видим в этом случае, что у нас здесь столько записей, что они полностью заполняют окно. Вот, при этом у нас максимальные и минимальные значения видны и те, которые там промежуточные, мы видим тоже. И вот мы здесь видим, что у нас конструкт появляется. То есть у нас вверху там значения свойств, которые больше всего характерны для этого класса, внизу те, которые больше в наибольшей степени не характерны. И они образуют собой противоположность смысловую, эти свойства.

Значит, я могу вам сказать, что информационные портреты классов, я их получил очень давно, реализовал. Даже вы, наверное, не представляете, насколько давно. Сейчас я вам покажу, насколько давно.

Значит, вот акт внедрения, акт о внедрении. Вот 87 год, Кубанский аэрокосмический центр. Я - главный конструктор в этом центре. И решаю задачу в среде системы Vega M, которую я разработал. Потом, когда через 13 лет появился Excel, то все ахнули, насколько похожа система Vega или насколько похож Excel на систему Vega M. Вот. Но система Vega имела некоторые преимущества даже вот перед современным Excel. Там вся графика была, всё то же самое, что сейчас в Excel. Таблицы можно обрабатывать было, но удобнее, чем сейчас в Excel, я скажу.

И вот смотрите, ребята, здесь видите, написано: позитивные, негативные информационные портреты 30 социальных категорий, с определением информативности признаков для распознавания социальных типов респондентов, обобщённые характеристики информативности признаков для набора такого, выбора такого минимального набора признаков, который содержит максимум информации. Вот. И позитивно-негативные информационные портреты. 87 год. Вот мы сейчас в системе этой видим эти позитивные, негативные информационные портреты классов. То есть любой класс выбираем и видим, какие позитивный портрет - это признаки в порядке убывания количества информации об этих признаках, о принадлежности объектов с такими признаками к этому классу. Это позитивный портрет. А негативный - это наоборот, наличие этих признаков не характерно, препятствует принадлежности к этому классу.

Мы здесь видим наличие, последний признак, видите, наличие проводов - нет. А первый признак - наличие проводов, провода есть. Кнопки есть, а здесь кнопок нет. То есть мы видим, что это конструкт признаков.

Я вам, ребята, рассказывал про конструкты, нет? Напомните мне, пожалуйста.

Да.

Рассказывал, да? Что это система противоположных по смыслу полюсов, образует одно единое понятие, имеющее полюса и промежуточный спектр, да, понятий. Вот. Ну вот мы видим, что эта система это делает. Итак, мы можем для любого абсолютно класса посмотреть конструкт.

Вот что характерно для монитора? Форма квадратная, наличие экрана, проводов. Вот. И что наиболее не характерно? Ну, наоборот, что экрана нет, проводов нет, кнопок нет. Вот. Думаю, что форма квадратная, студенты. Ну ладно.

3.4. SWOT-анализ

Следующий вопрос, связанный с исследованием модели - это SWOT-анализ. Что такое SWOT-анализ? Это то же самое, что позитивный, негативный информационные портреты, но они показаны отдельно друг от друга. Позитивный портрет показан слева, негативный справа. Вот. То есть такая форма, она является стандартной для экономистов. И в ней мы видим вверху, в верхнем окошке, целевое состояние фирмы, например, объекта управления, моделирования. Слева видим значения факторов, которые способствуют переходу объекта моделирования в это целевое состояние, справа - которые препятствуют. Нагрузки здесь вот видим, в какой степени способствуют, в какой препятствуют. И факторы эти все делятся на внутренние и внешние. Внутренние факторы находятся во власти руководства, оно может менять эти значения этих факторов. А внешние факторы - руководство не может на них влиять. Вот этим они отличаются. Это критерий, который позволяет отличить факторы управления от факторов окружающей среды.

Вот. То есть SWOT-анализ - это, в общем-то, та же самая информация, как в позитивных и негативных информационных портретах, но в такой вот форме, более традиционной, наверное. То есть, по крайней мере, экономисты изучают. И вот диаграмма соответствующая. Ну я так представлял себе, ну как это изобразить? Ну вот изобразил таким образом, что какие значения свойств или факторов более или в большей степени или в меньшей степени характерны или не характерны для объектов данного класса. Ну мы видим, что наличие проводов, кнопок, материал пластмассовый является очень характерными для элементов компьютеров. А, значит, наиболее нехарактерными является то, что проводов нет, и наличие кнопок тоже нет. Ну, противоположные, в общем, по смыслу свойства. То есть система вполне это может выявить эти особенности.

3.5. Анализ ценности признаков (Информационные портреты признаков)

Следующий момент, что мы можем узнать об объекте моделирования? Мы можем узнать, какие свойства этого объекта, какое влияние оказывают на принадлежность объекта с этими свойствами к различным классам.

Вот. Ну, допустим, чёрный, цвет чёрный или вот. О чём это говорит? Цвет чёрный. Значение свойства, значение, свойство цвет, а значение чёрное. Что это элемент компьютера, монитор, мышка, телефон, средства связи. И очень не похоже на то, чтобы это был спорт-инвентарь, мяч, мебель, аксессуар, сумка, клавиатура. Клавиатуры почему-то чёрные там не встретилось, что ли, я не знаю, стул. Это так по исходным данным.

3.6. Когнитивные диаграммы и нейросетевая визуализация

Смотрим теперь следующий момент, который мы можем исследовать - это вывод информации о том, что у нас представляют собой образ класса. Ну вот мы выводили уже в форме информационного портрета и в форме SWOT-диаграммы. А теперь визуализируем информацию о классе в той форме, которая традиционна для изображения нейрона. Поскольку модель допускает нейросетевую интерпретацию, то можно сказать так, что классу соответствует нейрон, и свойства его - это рецепторы, а значения свойств - это значения на этих рецепторах, которые фиксируются. И мы видим, что есть рецепторы, которые активируют нейрон красные, и которые тормозят этот нейрон синие. И они вот так расположены. То есть не очень много есть рецепторов, которые сильно влияют. И в основном остальное количество, в общем-то, мало влияет. Буквально несколько рецепторов оказывают сильное влияние, там, ну, три, четыре, там, пять, шесть, вот максимум. Остальные особой роли не играют.

И мы можем вывести все рецепторы, все нейроны вместе, в форме нейронной сети. Все нейроны с их рецепторами можем вывести вместе в форме нейронной сети.

Здесь мы что видим? Мы видим, что вот эта вся визуализация составляет примерно 60% наиболее значимых связей между рецепторами и нейронами. Красное - это положительная связь, синяя - отрицательная. Модуль связи - толщина линии. Толщина линии. И здесь как получена эта диаграмма одного слоя нейронной сети? Мы взяли, создали базу, в которой все связи отражены, вообще все, какие есть в модели. После этого эти связи все рассортированы в порядке убывания по модулю. То есть независимо от того, какой знак связи, по модулю в порядке убывания. И потом визуализированы из них те, которые наибольшее влияние оказывают. Причём визуализация наиболее сильных связей слева идёт, а потом всё меньше и меньше, слабее и слабее. И каждая связь характеризуется тем, что эта связь между определённым нейроном и определённым рецептором. Определённым значением даже снимаемым с рецептора. И вот так изображены все эти связи. Ну, можно, конечно, задать параметр, чтобы было больше 60%. Давайте попробуем. Вот у нас нейронов 16 отображено, их 14 всего. Мы разрешили до 16 визуализировать. А вот рецепторов у нас 50, а визуализировано тоже 16. Ну давайте зададим, допустим, ну, те же самые, что тут у нас? 50. И посмотрим, как это выглядит. Тогда у нас связи получится больше. Это выглядит следующим образом. И уже названия рецепторов не совсем нечитабельны практически, но коды ещё читабельны. То есть их такое количество, что всё-таки они здесь, хотя и чуть-чуть накладываются друг на друга, но ещё сохраняется читабельность.

Ещё можно, да, сейчас у нас 100% связей отображено, видите? То есть это полностью вся модель отображена, которая создана. Ну что можно сказать об этой модели? Так вот, взглянув на неё невооружённым взглядом, сразу бросается в глаза, что вот левые классы, они слева которые, они соединены с рецепторами толстыми линиями, а чем правее, тем линии тоньше. И, значит, это что означает? Что эти вот нейроны, соответствующие левым классам, они жёстко обусловлены значениями рецепторов. То есть есть влияющие сильно положительно, сильно отрицательно. То есть там очень определённость высокая. А вот самые правые, чем правее, тем меньше определённость, и самые правые нейроны, они обусловлены значениями рецепторов очень слабо. Вот это можно сказать, что, по-видимому, вариабельность свойств внутри этих вот рецепторов, которые сильно обусловлены. В общем так, они сильнее отличаются от остальных. И их легче идентифицировать.

3.7. Кластерно-конструктивный анализ

Теперь посмотрим на результаты кластерно-конструктивного анализа. Сначала нужно их рассчитать. То есть нужно рассчитать сначала матрицу сходства, которая является основой для когнитивных диаграмм и дендрограмм.

Мы её рассчитали. Теперь можем посмотреть в форме когнитивной диаграммы плоской, двумерной, какие у нас связи между классами, сходства и различия классов. Значит, сходство изображается толстой, то есть изображается красной линией, а различие - синей. А степень сходства различия изображается толщиной линии. Вот такие диаграммы могут быть нечитабельны. Поэтому есть средства...

(Телефонный разговор Георгия Васильевича с Евгением Вениаминовичем о смерти Валерия Ивановича и публикации статьи)

Вот. Значит, сейчас я вам покажу другой пример, другой диаграммы, где покажу, как настраивать, чтобы она была визуально, так сказать, можно было её воспринять. И показываю вам визуализацию той же самой матрицы сходства, ну не совсем только одной матрицы сходства, а ещё некоторый анализ этой матрицы сходства в виде когнитивной агломеративной дендрограммы. То есть дендрограмма - это иерархическая классификация. Значит, как она организована? Значит, есть много методов кластеризации, которые реализованы в системе Статистика, SPSS и других системах статистических. В системе Эйдос реализован алгоритм кластеризации, который я предложил. Есть статья соответствующая, где это я сделал. Значит, оно отличается чем от этих алгоритмов, которые описываются в литературе, реализованы в системах? Тем, что модель пересчитывается. То есть это этот алгоритм, который я предложил, он более трудоёмкий в вычислительном отношении. Я сейчас в двух словах его опишу. Сначала находятся по матрице сходства два наиболее похожих класса и объединяются в один класс. Потом в матрице исходных данных, то есть в таблице исходных данных, переставляются коды классов, к которым они относятся. То есть раньше считалось, что они относятся к двум классам, эти классы, объекты, объекта обучающей выборки относятся к этим двум классам. А теперь все объекты обучающей выборки, которые относились к первому, ко второму классу, которые в кластер объединены, у них проставляется, что они относятся к одному классу, который является вот этим кластером. А те классы из модели убираются, которые объединены, они убираются, а один обобщающий вставляется. И потом заново пересчитывается модель так, как она создавалась с первый раз. Только все объекты, которые относились к этим двум классам, которые объединились, они теперь относятся к этому кластеру теперь, к объекту, который соответствует кластеру. Потом опять рассчитывается матрица сходства, определяется наиболее похожий класс. И причём этот кластер уже тоже как класс рассматривается. И следующие объекты наиболее похожие объединяются, два класса объединяются. В частности, этим классом может быть и кластер. Вот. И продолжается этот процесс до тех пор, пока остаётся два класса в модели. И они просто визуализируются уже. Вот это агломеративная кластеризация. Агломеративная - то есть объединительная, то есть объединение происходит классов. Вот. А в классическом кластерном анализе, там что делается? Если два класса наиболее похожи по матрице сходства, то они объединяются в один кластер. А потом в этой матрице сходства все коэффициенты, которые отражают сходство различия этих классов с другими классами, заменяются на средние для этих двух классов. Вот один класс похож настолько-то на какой-то класс, а другой настолько-то похож на такой-то класс. Ставится среднее. Ну то есть пересчитывается матрица сходства. Это очень такая рискованная операция, которая имеет смысл и корректно только при определённых условиях. Ну, при том, что вес объектов, которые относятся к этим двум классам, примерно одинаковый. Что выборка такова исходных данных, что примерно одинаковое количество объектов по этим классам и так далее, и так далее. То есть это очень много предположений возникает. И в корректности этой операции. Ну я могу вам сказать, допустим, вот объединяется два класса: стол, допустим, и мебель. Ну представьте себе, что в мебели у нас 1000, например, примеров объектов, а стол - пять примеров объектов. Вот мы их объединяем в общий кластер, да? А потом берём матрицу сходства, делим, средние пишем уровни сходства для этих двух классов. Какое у нас основание писать среднее? Там в 20 раз больше или там в 200 раз больше объектов первого класса, чем второго. То есть у нас явно должно быть смещено это эта величина, не средняя должна быть, а она должна быть сильно смещена в сторону того класса, где уровень сходства для того класса, которых больше примеров. Вот. Но это ж ещё не всё, они же эти примеры могут быть типичные, нетипичные, то есть они в разной степени сходны с другими классами и так далее. То есть на самом деле там вообще не поймёшь, как это усреднять. Ну, по крайней мере, не так, как это делается, то уж точно, не поровну. Ну поровну - это такое грубое очень упрощение, понимаете, очень такое жёсткое допущение, что это корректно. То есть я думаю, что это некорректно, вообще говоря. Вот. Конечно, быстро получается, но неправильно. То есть у меня есть опыт применения стандартного кластерного анализа. Меня просили там с разных кафедр посчитать эти дендрограммы. Я считал, приводил их там. И меня поразило, что происходит потом после этого. После этого, ну я там посчитал, я так прикинул, что где-то там больше 100 штук получится, если все их посчитать. Ну я не стал их считать все, если честно. Посчитал, сколько было не лень, как говорится, штук 30, наверное, там 40, вот так где-то. Вот, я отдал заведующему кафедрой. И он взял и стал их листать и говорить, что это вот комментировать, что это вот какая-то не очень, там, это тоже вообще странная какая-то, полный отстой. И вот так вот комментировал, комментировал. И потом говорит: "А это вот вроде нормальная". Раз её отложил. И у меня возник очень большой вопрос по этому поводу, что является ли кластерный анализ классический методом исследования? Или скорее можно считать, что он является способом подгонки результатов? То есть фактически, о чём идёт речь? О том, что... (показывает статью) ...это я вам показал статью по кластерному анализу своему и послал ссылочку на неё. По сути дела, идёт речь о чём? Я ему говорю этому зав кафедрой: "А что бы вам не взять Photoshop там или что-нибудь Corel Draw там или Paint на худой конец? Или вообще MS Visio, например? И взять и просто нарисовать эту дендрограмму, какая вам нравится? Потому что всё равно сводится к тому, чтобы выбирать, какая вам нравится из сотен вариантов, понимаете?" Вот. А почему он выбирал те, которые ему нравятся? Знаете, почему? Вот ответьте на этот вопрос, пожалуйста. Кто-нибудь может догадаться, почему он выбирал те, которые ему нравятся? И откладывал те, которые ему не нравятся. Не соответствуют его интуитивным представлениям. Давайте, отвечайте на вопрос. Я к вам обращаюсь, студенты. Георгий Александрович, он, конечно, ответит, но я его не спрашиваю, он это преподаватель же, правильно? А вот давайте студенты пусть отвечают на вопрос. Вот, допустим, Михаил, староста ж всё-таки, давай.

Я не знаю

Ну как это не знаешь? Ну скажет: "Вот это мне нравится, это не нравится". Он эксперт в этой области, он хорошо в ней разбирается содержательно, глубоко. Он этим занимается всю жизнь. Что означает, когда мы из его уст слышим слова "не нравится мне это"? Что это значит? Это значит, что эта дендрограмма фонарная, понимаете, что она вообще отстойная, что она не отражает реальные области. Если эксперт смотрит и говорит, что это вообще ерунда какая-то, и выбрасывает её, понимаете, на черновики. Что это значит? Это означает, что система, которая делает сотни таких дендрограмм, что-то эта система тоже отстойная, понимаете? Вот о что это значит. То есть она не отражает реальных закономерностей предметной области. Вот что это значит. То есть это вот эти методы, которые я в этой статье раскритиковал кластерного анализа, они некорректны. Там евклидово расстояние используется для неортонормированных пространств, там мера сходства вот используется, как это вот евклидово расстояние. Там используются вот эти вот пересчёт матрицы сходства. Непонятно, с какого перепуга там средние берутся. То есть там масса вопросов возникает, понимаете? Потом там исследуется обычно матрица абсолютных частот, которая в разных единицах измерения, в разных типах шкал. Как можно их совместно обрабатывать, вообще непонятно. Это ж некорректно. Это то же самое, что там один студент имеет рост метр 73, а другой 183 см там. И взять и сложить 183 прибавить ещё метр 73. Там или там, допустим, там... Ну получится там 187 там или 188 там и 43 там что-нибудь там сотых. Глупость, понимаете? То есть складывать, умножать величины в разных единицах измерения - это глупость полнейшая. За это двойки ставят в третьем классе, понимаете? Когда мы берём расстояние, складываем сантиметры с метрами, складываем. Одна книжка была размером 30 см, а другая книжка была 0,4 м. Мы взяли, сложили их размеры и посмотрели, какая будет, какое место они будут занимать на столе. 30 + 0,4 получилось 30 + 30,4. Ну это ж глупость полнейшая, понимаете? Чтобы это складывать, нужно перевести в одну единицу измерения. Они это делают в системе Статистика? Как вы думаете? Я вообще в шоке, ребята, но нет, они этого не делают, понимаете? Они этого не делают. Они это оставляют на усмотрение пользователя. Хочешь складывать сантиметры с километрами, хочешь там что хочешь, литры там с килограммами. Твои проблемы, понимаешь? А я понимаю, что пользователь не будет переводить это в одну единицу измерения всё. Не будет, потому что у него нет такой возможности. Поэтому в системе Эйдос всё это переводится в одну единицу измерения. В качестве которой выбран наиболее универсальный, на мой взгляд, подход - это количество информации. То есть я рассматриваю количество информации, которое содержится в признаке о принадлежности объекта с этим признаком к этому классу. И мне без разницы, в каких единицах измерения этот признак измерен, понимаете? Какие бы ни было измерено, я вычисляю количество информации. И потом дальше можно корректно с этим с этими всеми данными работать, потому что они уже в одной единице измерения. То есть решена важнейшая проблема - проблема обеспечения сопоставимости обработки данных модели.

3.8. Значимость классов и признаков

Смотрим теперь значимость градаций описательных шкал, то есть признаков. Насколько они являются ценными для решения задачи идентификации. Ну что мы видим, ребята? Мы видим, что если взять все признаки и расположить их в порядке убывания их ценности, так что сначала будут идти наиболее ценные, потом менее ценные, то ценность будет расти суммарная сначала быстро, а потом всё медленнее и медленнее. Вот в данном случае мы видим, что 23% наиболее значимых признаков обеспечивают 50% суммарной значимости. То есть, видите, четверть там, пятая, почти что часть признаков обеспечивает половину суммарной значимости. А половина наиболее значимых признаков обеспечивает 80% суммарной значимости модели F3. А здесь даже ещё большее различие. 18% наиболее значимых признаков 50% значимости обеспечивают.

Вот. Ну это значит, что можно вот эти признаки, которые не очень ценные, просто удалить. Но признаки мы удалить не можем, а мы можем удалить только шкалу. Посмотрим на выходные формы, на сами матрицы моделей.

Вот здесь мы видим, ребята, какие признаки наиболее ценные. Наличие проводов есть, нет - это содержит 5% суммарной значимости. Наличие кнопок есть, нет - ещё 5%. Материал пластмасса - 4%, форма круглая, форма и так далее. Значит, ясно, что шкалы наличие проводов, наличие кнопок очень ценными являются. Ну давайте теперь посмотрим по шкалам такую же форму.

Ну там усреднённая информация по ценности признаков, она усреднена по шкалам. Ну здесь не рисуется, просто форма, то есть просто база данных формируется. И мы эти базы данных посмотрим. И вот мы видим, что наиболее ценными являются шкала наличие проводов, наличие кнопок, наличие экрана. Вот наличие проводов - это 20% суммарной значимости. Ещё кнопок - ещё 20%. То есть получается что, что мы берём две шкалы и уже получаем 40% суммарной значимости. А если ещё берём третью, тогда получаем больше 50% суммарной значимости. А ещё шесть шкал, да? Шесть, они имеют значимость суммарную меньше 50%, а три первых - больше 50%, 54%. Ну если это хочется нарисовать, то проблем никаких нет, надо нарисовать.

То же самое касается классов. То есть мы можем посмотреть на то, насколько у нас классы жёстко обусловлены значениями факторов. Есть классы, которые жёстко обусловлены, есть, которые менее жёстко обусловлены. Вот. То есть мы всё это видим.

Ну вот, допустим, в этой модели мы видим, что четверть классов обеспечивает половину суммарной значимости. Что это за классы, мы сейчас посмотрим.

Ну ладно. В общем, короче говоря, видно, какая ценность этих классов. То есть насколько они жёстко обусловлены входящими, то есть теми свойствами, которые для них характерны.

**4. Задание для студентов**

Так, значит, сейчас мы пришли к тому, что мы закончили вопрос об исследовании предметной области, ребята. И мы сейчас должны приступить к рассмотрению следующего вопроса, который состоит в том, что мы начинаем решать собственные задачи. Значит, у нас обучение предусматривает выполнение определённых заданий. То есть мы не просто вот слушаем там всё, смотрим на экран, а ещё должны что-то понять, усвоить из этого и потом решать задачи свои, используя вот эти навыки, которые получены у нас в процессе обучения.

И вот сейчас я вам даю задание, ребята. Вот я вам послал ссылочку в чат на это задание. Бывают иногда такие курсы, где задание на каждую, на каждое лабораторное занятие или на каждое лекционное занятие предусмотрено задание. А я предусмотрел комплексное задание, которое проверяет всю совокупность умений и навыков, знаний, умений и навыков. И это комплексное задание заключается в том, что вы должны разработать собственное приложение.

Так, теперь давайте мне сейчас подтверждение вашего присутствия. Значит, в чат пишите плюсики, кто меня сейчас слушает. Потому что я сейчас даю задание, и мне важно, чтобы все меня хорошо слушали и слышали.

Ну, Михаил, ты, значит, тем, кто отсутствует сейчас на занятии, передай эту ссылочку и скажи, что оценка будет ставиться жёстко по тому алгоритму, который описан в этой инструкции и задании. Сейчас я этот алгоритм вам приведу. Вот.

Так, ну я сколько ж вас тут человек? У нас семь человек здесь слушателей сейчас, учащихся. Не бог весть. Но плюсики поставили трое. А ещё четверо что делают сейчас? Вообще положили телефончик там на стол и пошли себе спать, да? Или смотреть телевизор, или чем они там занимаются?

Ну, Михаил, выясни, чем остальные занимаются сейчас. Почему они плюсики не пишут? Ну на Михаил, выясни, чем остальные занимаются сейчас. Почему они плюсики не пишут? Даже из тех, кто присутствует. Три человека на занятии, ребята, это вообще отстой полный.

А я даю задание. Говорю об оценке результатов обучения. И меня никто не слушает, потому что нету их просто там, понимаете? Связь указана, что вроде как связь есть, а фактически их нету.

Да, это грустно всё видеть. Соответствующее будущее вас ждёт, те, кто так учился. Ну те, кто так учился, даже этого не слышат, потому что они так учатся, что даже и не слышат, что говорит преподаватель.

Значит, Михаил, доведи до всех 100% студентов вот эту ссылочку. Скажи, вот есть ссылочка на задание и инструкцию, как её выполнять это задание. И последующие занятия посвящены именно этому. И там есть и система оценки, критерии оценки. И по этим критериям будет оцениваться, вы будете оцениваться, и вам будет выставляться оценка.

То есть, смотрите, вы должны найти данные. Я сейчас вам описал, показал вам много приложений в процессе обучения. Показал вам лабораторные работы, в которых описано, показано, и мы изучали эти работы подробнейшим образом, вам всё объяснял. И те, кто ходили на занятия, они это могли слышать. Как решать задачи на основе внешних данных, представленных в экселевской форме табличной. Данные и числовые, и текстовые. Как решать задачи по интеллектуальной обработке текстов? Как решать реальные задачи по оценке объектов недвижимости или другие задачи, связанные с уже реальной обработкой информации из таблиц экселевских. И как решать задачи, связанные с обработкой изображений. Я вам показал не все подходы, которые есть в системе Эйдос, а только один из четырёх или из пяти подходов, где показано было, как мы можем исследовать изображения, анализируя их спектры и формировать обобщённые спектры классов, сравнивать конкретные спектры конкретных изображений с обобщёнными спектрами классов и так далее, и так далее.

Я вам всё это показал, рассказал подробно. Вот. То есть вы должны сейчас уже обладать знаниями, умениями и навыками для создания собственных приложений. Если у вас какие-то возникают с этим проблемы, то на занятиях вот этих лабораторных, вы можете меня спрашивать всё. Как раз на этих занятиях мы должны вам давать какую-то поддержку, методическую помощь в решении самостоятельных задач. То есть вы сейчас должны что сделать после того, как вот это занятие закончится? Вы должны зарегистрироваться в ResearchGate и РИНЦ. Вот. В РИНЦ зарегистрироваться также в системе Science Index, получить спин-код и заключить договор на размещение непериодических публикаций в базах данных РИНЦ. Тогда вы сможете этот пункт выполнить соответствующий РИНЦ, разместить, разместить. Вот. В ResearchGate вы должны зарегистрироваться, чтобы там вы могли разместить своё описание своего приложения. И вы должны сделать это приложение, сделать его описание. Тогда вы сможете зарегистрировать его в Эйдос-облаке. Вот, это FTP-сервер системы Эйдос. Если вы хотя бы это сделаете, вы уже два балла не получите. Если же вы ничего этого не сделаете из того, что описано выше, тогда будет оценка по результатам аттестации, которая будет низкая, низкая оценка по результатам аттестации. Ну, скажем так, два балла я не поставлю, но, наверное, там три в лучшем случае, вот так вот, в общем. Понятно? То есть это будет очень тогда для вас нежелательно. Ну я так понимаю, что вам всё равно, потому что вы ж на занятия не ходите, какая вам разница? Поэтому вполне нормально.

Так, значит, смотрим сейчас, кто там не поставил плюсик. Елисеева поставила, Тигров и Маргар и Чухарева поставили, да? Вот. Ацулов не поставил. Вот я ему написал вопрос: "Ты где?" Пока он на него не среагировал. Александра, ты слышал, что я сейчас говорил про то, что вот тут присутствуют, вот всё это вот? Возникли проблемы с интернетом. И у остальных что ли возникли проблемы с интернетом? Угу. Ну ладно.

**5. Следующий вопрос и завершение:**

Тогда мы начинаем следующий вопрос. Теперь мы должны выполнить пункт четыре. Пункт четыре в чём заключается? В том, что мы должны выбрать тему приложения, которое будем разрабатывать. Для этого мы должны посмотреть на приложения, которые уже созданы. Ваше приложение не должно повторять те приложения, которые уже созданы. Но если приложение создано давно, то, в принципе, можно там спросить меня и, может быть, и можно даже и сделать опять же это же самое приложение, но на другом совершенно уровне качества, используя современные возможности и современные шаблоны для описания этих приложений.

Вот. И потом ищем данные. Есть несколько способов это сделать. Простейший способ - это можно задать запрос "данные для машинного обучения", "наборы данных для машинного обучения". По этому запросу находится огромное количество сайтов, на которых есть данные для машинного обучения. Можете там выбрать те, которые вам подходят, интересны для вас. Вот. Также наборы данных для машинного обучения есть на сайтах UCI. Здесь их очень много. Но надо, я сказал уже, брать только те можно, которые ещё не были реализованы. А реализовано было довольно много. Штук 10, наверное. Первое приложение на основе данных репозитория UCI я решил в 2003 году. То есть сколько лет назад? Аж 17 лет назад. Вот она статья того времени. Вот. И потом ещё возвращался к этой теме не раз, а именно семь раз. Именно с репозиторием UCI. И другие тоже использовал источники данных для машинного обучения. Вот. Более поздних статей у меня нет по этой, именно вот с этим репозиторием.

То есть репозиторий UCI является, по-видимому, наиболее древним порталом, на котором находятся данные для машинного обучения. То есть были небольшие сайты, где что-то там какие-то фрагменты были, но несопоставимы были с этим репозиторием UCI. То есть там, конечно, гораздо больше база данных этих вот для машинного обучения датасетов. Вот. Потом не так давно, ну, сравнительно, если сравнить этот 17 лет, ну даже не вот это я нашёл 18 лет назад, а он-то был и раньше, ну лет 20 уже, наверное, он существует. А вот Kaggle, он не так давно, сравнительно, ну может лет пять там, не знаю, там историю можно почитать. Возник портал, который те, кто его создал, они, видимо, как-то поняли, что просто разместить данные - это ещё недостаточно, нужно как-то заинтересовать этих вот пользователей, чтобы они какую-то дополнительную мотивацию получили.

Значит, и создали дополнительную мотивацию. В чём она заключается? В том, что на сайте Kaggle не просто размещены огромное количество различных наборов данных, которые, по-моему, гораздо больше, чем на UCI. Но что очень существенно, там есть так называемые конкурсы, открытые конкурсы. И есть завершённые конкурсы. Так вот, на конкурсах есть призовой фонд. И те, кто побеждает, занимают три призовые места, им даётся этот призовой фонд. И он может составлять там от 3-5 тысяч долларов до 100 тысяч долларов, я видел призовой фонд. То есть, занимаешь первое место, получаешь половину. Ещё два получают 1/3 и 2/3 от оставшегося. Примерно так. Ну я точно не помню, но где-то вот примерно такое разделение. Ну это очень вообще-то серьёзно. И есть люди, значит, ну, у всех, значит, там надо, чтобы получить данные, надо зарегистрироваться на сайте Kaggle. То есть там данных побольше, чем на UCI. Значит, там есть уже завершённые конкурсы, текущие конкурсы есть. И есть песочница, которая вообще не является конкурсом, а просто можно потренироваться. И вот есть люди, я там зарегистрировался в определённое время назад. И по своему статусу вот я являюсь новичком. Я решил, принял участие в двух соревнованиях. В одном получил результат... Ну я не знаю, что здесь, как это понимать. Вот. Ну, в общем, в одном написано "лучше 85%", в другом "91". Я лучше или они лучше, я не совсем понимаю. Вот. Ну, в общем, короче говоря, там есть люди, которые постоянно занимают эти призовые места. Прямо регулярно, понимаете? Они там как работают. Вот. Что, в общем-то, интересно. То есть есть люди, которые обладают такими навыками, технологиями, которые позволяют им решать задачи очень хорошо.

Значит, на следующем занятии, ребята, я хотел бы, чтобы вы выбрали тему уже и почитали эту инструкцию.