

УДК 004.89

DOI: [10.26102/2310-6018/2019.27.4.006](https://doi.org/10.26102/2310-6018/2019.27.4.006)

ПРИМЕНЕНИЕ ИНСТРУМЕНТОВ DATA MINING ДЛЯ АНАЛИЗА И ПРОГНОЗИРОВАНИЯ УДОВЛЕТВОРЕННОСТИ ОБУЧЕНИЕМ

П.Б. Скрипко¹, В.С. Дунин², Е.А. Клейменов³

*ФГКОУ ВО Дальневосточный юридический институт министерства внутренних дел
Российской Федерации, Хабаровск, Россия*

¹e-mail: skripkop@yandex.ru

²e-mail: dvs_82@mail.ru

³e-mail: yevgeni-k1@yandex.ru

Резюме: В статье рассматриваются особенности и результаты применения инструментов интеллектуального анализа данных – приложения для построения деревьев решений See5, с помощью которого на основе данных социологического опроса определяется ряд параметров, наиболее значимых для оценки удовлетворенности обучением. Наряду с особенностями применения программного инструмента анализа данных в работе показан порядок подготовки данных для загрузки и дальнейшего анализа. При этом представлены не только форматы файлов, загружаемых в приложение, но и подходы к очистке и агрегированию исходных данных, в том числе и в целях снижения их размерности. Интерпретация результатов построения деревьев решений – правил классификации, полученных в ходе обработки – направлена в том числе на выбор оптимальных настроек обработки исходных данных. Построенный с использованием данных социологического опроса классификатор показал, что существенное влияние на классификацию оказывают такие параметры, как атмосфера межличностных отношений, признание достижений обучаемого, престиж образовательной организации. Также в работе приведена процедура оценки прогнозных значений удовлетворенности обучаемых на основе ввода предполагаемых значений классификационных признаков.

Ключевые слова: инструментарий интеллектуального анализа данных, деревья решений, удовлетворенность обучаемых, классификаторы, решающие правила, прогнозирование удовлетворенности обучением.

Для цитирования: Скрипко П.Б., Дунин В.С., Клейменов Е.А. Применение инструментов Data Mining для анализа и прогнозирования удовлетворенности обучением. *Моделирование, оптимизация и информационные технологии*. 2019;7(4). Доступно по: https://moit.vivt.ru/wp-content/uploads/2019/11/SkripkoSoavtors_4_19_1.pdf
DOI: 10.26102/2310-6018/2019.27.4.006.

APPLICATION OF DATA MINING TOOLS FOR ANALYSIS AND FORECASTING OF LEARNING SATISFACTION

P.B. Skripko¹, V.S. Dunin², E.A. Kleimenov³

*Far Eastern Law Institute of the Ministry of the Interior of the Russian Federation,
Khabarovsk, Russia*

Abstract: The article discusses the features and results of using data mining tools, an application for building See5 decision trees, with which a number of parameters are determined on the basis of sociological survey data that are most important for assessing student satisfaction. Along with the features of using a software tool for data analysis in the work, the procedure for

preparing data for downloading and further analysis is shown. At the same time, not only file formats loaded into the application are presented, but also approaches to cleaning and aggregating source data, including in order to reduce their dimensionality. Interpretation of the results of constructing decision trees - classification rules obtained during processing, is aimed, inter alia, at choosing the optimal settings for processing the source data. The classifier constructed using the data of a sociological survey showed that such parameters as the atmosphere of interpersonal relations, recognition of the achievements of the student, and the prestige of the educational organization have a significant impact on the classification. The paper also describes the procedure for assessing the predicted values of student satisfaction based on the input of the estimated values of the classification features.

Keywords: data mining tools, decision trees, decision rules, classifiers, forecasting student satisfaction.

For citation: Skripko P.B., Dunin V.S., Kleimenov E.A. Application of Data Mining tools for analysis and forecasting of learning satisfaction. *Modeling, optimization and information technology*. 2019;7(4). Available by:

https://moit.vivt.ru/wp-content/uploads/2019/11/SkripkoSoavtors_4_19_1.pdf

DOI: 10.26102/2310-6018/2019.27.4.006 (In Russ.).

Введение

Сегодня одной из актуальных задач совершенствования образовательных программ высшего образования является повышение уровня удовлетворенности обучением. Для решения такого рода задач проводятся специальные социологические исследования, в ходе которых на основе данных опроса обучаемых формулируются выводы как об удовлетворенности обучением в целом, так и удовлетворенности отдельными элементами образовательных программ [1]. Это соотношение объема теоретических и практических занятий, качество и доступность библиотечного фонда и электронной информационно-образовательной среды, расписание занятий, порядок и условия промежуточной и итоговой аттестации, прохождения практик и др. Полученные в ходе опроса данные обрабатываются и анализируются, группируются по различным признакам и представляются для дальнейшего изучения и использования в целях совершенствования деятельности образовательной организации, повышения качества образовательных программ, улучшения условий обучения [2]. Однако методы обработки, анализа и представления данных об удовлетворенности обучением основаны на применении predetermined запросов и построении графиков и диаграмм, иллюстрирующих различные количественные соотношения ответов обучаемых [3]. При этом для обоснованного ответа на вопрос о том, какие параметры образовательной деятельности и в какой степени оказывают влияние на удовлетворенность обучением в целом и отдельными составляющими учебного процесса требуется существенный объем данных, исследования основанные на фокус-группах не обеспечивают необходимого количества наблюдений. Для ответа на этот и подобные вопросы необходимо провести более глубокий анализ результатов опроса в целях обнаружения в них некоторой «скрытой» информации, для извлечения которой сегодня доступны методы «добычи данных» – методы Data Mining.

Программные системы (приложения) для интеллектуального анализа данных становятся все более распространенным инструментом для понимания и решения различных задач и могут эффективно применяться к обработке больших объемов данных [4, 5]. К таким задачам в образовательной деятельности можно отнести установление основных причин неудовлетворенности обучаемых получаемыми образовательными

услугами, определение условий, способствующих повышению качества обучения, а также предсказание изменений в удовлетворенности обучаемых при введении новых образовательных программ или изменении ключевых параметров уже реализуемых [6, 7]. Результаты, полученные в ходе их решения, несомненно окажут существенную поддержку организационно-управленческой составляющей деятельности образовательной организации и, в конечном счете, позволят повысить качество высшего образования.

Материалы и методы

В качестве инструмента для решения задачи классификации результатов опроса обучаемых об удовлетворенности обучением выбрано приложение для интеллектуального анализа данных See5/C5.0, разработанное компанией Rulequest (<https://www.rulequest.com>). Последняя версия (release 2.11a) этой программы датирована апрелем 2019 года, что говорит об ее актуальности и распространенности. See5 предназначена для анализа больших наборов данных, содержащих до сотни тысяч описаний объектов наблюдений, описываемых до ста числовыми или номинальными параметрами. Результат работы этого приложения выражается в виде деревьев решений и множества if-then правил. На наш взгляд данное приложение является достаточно простым и доступным инструментальным средством для классификации на основе деревьев решений и не требует от пользователя специфических знаний в области прикладной статистики [8].

Рассмотрим особенности применения данного приложения, разделив процесс анализа на три классических для данной области этапа: подготовка данных, построение деревьев решений (логических правил) и интерпретация результатов (прогнозирование) [9].

При простоте использования и дальнейшей интерпретации результатов работы данного приложения достаточно затратным с точки зрения времени является этап подготовки данных для загрузки в систему. Так, в нашем случае (анализ результатов опроса обучаемых об удовлетворенности образованием) опрос проводился с использованием ресурсов электронной информационно-образовательной среды института, доступ к которой осуществлялся посредством сети Интернет. Результаты опроса были выгружены в текстовом формате (UTF-8) с разделителями – в виде файла CSV, над которым с помощью табличного процессора Microsoft Excel был выполнен ряд преобразований в целях снижения размерности (числа переменных, параметров) и формализации признаков классификации. Наряду со снижением размерности исходных данных, отмеченные ниже преобразования обусловлены также и невозможностью повторения опроса.

В ходе указанных преобразований произведена замена свободных ответов (вариант ответа «Другое») и ввод на вопросы анкеты на значения «1» при наличии ответа в такой форме и на «0» соответственно при отсутствии.

Для открытых вопросов анкеты, ответы на которые предполагают указание в свободной форме нескольких мнений или предложений, как например, для вопроса «Какие дисциплины Вы хотели бы добавить в учебную программу ДВЮИ МВД России?», произведен подсчет количества ответов для каждого из подобных вопросов. Полученные значения использованы в качестве значений параметров.

Ответы на вопросы анкеты сгруппированы по направлениям деятельности обучаемых в институте с учетом специфики образовательной организации МВД России. Так в первую группу вошли ответы на вопросы, связанные с удовлетворенностью институтом в целом, получаемой специальностью; во вторую – оценка качества учебной деятельности, в том числе работы преподавателей и мнения обучаемых по повышению

качества учебы; в третью, наиболее крупную группу, вошли вопросы об условиях обучения с учетом следующих направлений: служебная деятельность (наряды, участие в службе по ООП), различные аспекты институтской жизни (межличностные отношения, организация культурно-массовых, спортивных мероприятий, работа кураторов, медицинское обслуживание, организация научной работы обучаемых), информатизация института, библиотека, возможности для самостоятельной работы; четвертая группа – это оценка организации практик и стажировок; пятая группа – оценка организации и проведения научной работы обучаемых; шестая группа – оценка отношения «обучаемый» - «преподаватель»; последняя, седьмая группа содержит характеристики объекта опроса – специальность, год обучения, комплекующий орган.

Ответы на вопросы, в которых обучаемых напрямую спрашивают об удовлетворенности тем или иным направлением деятельности института, выбраны в качестве составляющих целевого параметра. Данные параметры объединены в единый показатель удовлетворенности обучаемых, на основе которого проведена классификация. Таких ответов в анкете насчитывается 8, каждый из которых содержит значения от 1 до 5 (1 – полностью удовлетворен, 2 – скорее удовлетворен, 3 – скорее не удовлетворен, 4 – полностью не удовлетворен и 5 – затрудняюсь ответить). Процедура их объединения заключается в суммировании значений и дальнейшем разделении на интервалы для приведения целевого параметра к принятым значениям от 1 до 5.

В результате суммирования полученные значения целевого параметра находятся в диапазоне от 8 (обучаемый полностью удовлетворен по всем составляющим) до 40 (обучаемый затрудняется ответить). Этот диапазон разделен на 5 интервалов: 8-13 (полностью удовлетворен), 14-20 (скорее удовлетворен), 21-27 (скорее не удовлетворен), 28-34 (полностью не удовлетворен) и 35-40 (затрудняюсь ответить). В зависимости от попадания значения целевого параметра в указанные интервалы приведены в соответствие с принятыми оценками от 1 до 5 (от «полностью удовлетворен» до «затрудняюсь ответить»).

Аналогичная процедура объединения выполнена и для ответов на вопросы анкеты, в которых предлагается оценить удовлетворенность одновременно по нескольким аспектам деятельности института. Так, например, для ответов на вопрос анкеты «В жизни нашего Института имеется много различных сторон и аспектов, которые так или иначе затрагивают каждого курсанта (слушателя). Оцените, насколько Вы удовлетворены?» в результате такой операции получены значения в диапазоне от 14 (обучаемый полностью удовлетворен по всем направлениям) до 70 (обучаемый затрудняется ответить по всем направлениям). Полученный диапазон разделен на соответствующие 5 интервалов: 14-24 (полностью удовлетворен) 25-36 (скорее удовлетворен) 37-48 (скорее не удовлетворен) 49-60 (полностью не удовлетворен) 60-70 (затрудняюсь ответить). Каждому из полученных интервалов поставлена в соответствие оценка от 1 до 5 (от «полностью удовлетворен» до «затрудняюсь ответить»).

В ходе анализа результатов опроса обучаемых об удовлетворенности получаемым образованием отмечена высокая лояльность обучаемых к образовательной деятельности института – более 97 % опрошенных ответили на вопросы как «полностью удовлетворен» и «скорее удовлетворен», а оставшиеся 3 % приходятся на ответы «скорее не удовлетворен», «полностью не удовлетворен» и «затрудняюсь ответить». В этой связи было в целевом параметре учтены только два класса: 1 класс – «полностью удовлетворен», соответствует ответу на вопросы анкеты «полностью удовлетворен» и принимает значение «1»; 2 класс – это «частично удовлетворен», включает ответы «скорее удовлетворен», «скорее не удовлетворен», «полностью не удовлетворен», «затрудняюсь ответить» и принимает значение «2». Таким образом, ответы, относящиеся к 1 классу – «полностью удовлетворен», определяются удовлетворенностью обучаемых

всеми (или большей частью) направлений деятельности института, а ответы 2 класса – «частично удовлетворен» учитывают, что к ряду направлений у обучаемых имеются претензии.

В результате выполненных на данном этапе преобразований результатов опроса получен набор исходных данных из 300 случаев (ответов), описываемых 218 атрибутами. Атрибуты соответствуют ответам на вопросы анкеты.

Для загрузки в приложение See5 и дальнейшую обработку исходные данные должны быть представлены как минимум двумя файлами в текстовом формате UTF-8 – файлами имен переменных и данных. В файле имен переменных, имеющем расширение .names представлены названия используемых атрибутов и классов.

Фрагмент файла имен переменных для нашей задачи students.names представлен на Рисунке 1.

SATF.	the target attribute
SATF:	1, 2
A_1->1:	0, 1
• • • • •	
A_2:	continuous
A_3:	ignore
• • • • •	
A_7-> count:	continuous
• • • • •	
B_12:	2, 3, 4, 5
• • • • •	
C.2_30->1:	1, 2, 3, 4, 5
• • • • •	
F_50->1:	0, 1
• • • • •	
G_56:	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

Рисунок 1 – Фрагмент файла имен переменных
Figure 1 - A fragment of a file of variable names

Слева на Рисунке 1 указаны имена атрибутов, а после двоеточия – тип значений, который может принимать соответствующий атрибут. Так атрибут принимающий строго определенные значения является номинальным признаком (discrete attribute), а атрибут принимающий любые числовые значения – количественным признаком (continuous attribute).

Целевой признак SATF принимает два значения: 1 – в классе «полностью удовлетворен» и 2 – в классе «частично удовлетворен». Признаки A_1->1 (ответ «1. Советом родных, друзей, знакомых» на 1 вопрос анкеты «Чем Вы руководствовались, выбирая ДВЮИ МВД России?»), B_12 (ответ на 12 вопрос анкеты «Как бы Вы оценили качество образования, получаемого в ДВЮИ МВД России по шкале от 2 до 5?»), C.2_30->1 (ответ «1. Отношением со стороны руководства курса к Вам» на 30 вопрос анкеты «В жизни нашего Института имеется много различных сторон и аспектов, которые так или иначе затрагивают каждого курсанта (слушателя). Оцените, насколько Вы удовлетворены?»), F_50->1 (ответ «1. Уважение» на 50 вопрос анкеты «Как бы Вы охарактеризовали Ваше типичное отношение как курсанта и слушателя к преподавателям при проведении занятий?»), G_56 (ответ на 56 вопрос анкеты «Укажите регион, в котором находится Ваш комплектующий орган») относятся к номинальным, после двоеточия указаны значения, которые они могут принимать в исходном наборе

данных. Признак A_7->count (количество указанных ответов на вопрос анкеты «Если Вы в целом не удовлетворены обучением в ДВЮИ МВД России, то укажите причины этого») относится к количественным признакам и принимают различные числовые значения. Признак A_3 (ответ на вопрос анкеты «Удовлетворены ли Вы процедурой поступления в ДВЮИ МВД России?») входит в состав целевого признака и в ходе классификации не используется (ignore).

Файл данных является вторым входным набором данных необходимым для работы See5 и имеет расширение .data. Фрагмент файла students.data представлен на Рисунке 2.

```

2,0,0,0,1,0,1,1,1,0,0,1,3,2,2,3,0,0,0,0,1,0,0,0,0,0,1, ... 0,0,1,0,0,0,1,0,0,0,1,1,1,5,2
1,0,1,0,1,0,1,0,0,0,0,0,1,1,1,3,0,1,1,0,0,1,1,0,0,1,0, ... 0,0,1,0,0,0,1,0,0,1,2,2,1,5,2
1,0,1,1,1,0,1,0,0,0,1,0,1,1,1,3,1,1,1,0,1,0,0,0,0,0,0, ... 0,0,0,0,0,0,0,0,1,1,1,2,2,5,2
2,0,1,1,1,0,1,0,1,0,0,0,1,1,2,3,0,1,0,0,1,1,1,0,1,1,1, ... 1,0,0,0,0,0,0,0,0,1,1,1,3,5,2
2,1,1,1,0,0,0,0,0,0,0,0,2,2,2,3,0,0,1,0,0,1,0,0,0,1,1, ... 0,0,0,0,0,0,0,0,1,1,2,2,5,5,2
2,1,1,1,1,1,0,1,1,0,0,2,2,2,3,1,1,1,1,1,1,1,1,1,1,1, ... 1,0,1,0,0,0,1,0,0,1,1,1,5,5,2
1,1,0,0,0,1,0,0,0,0,1,0,1,2,1,5,1,0,1,0,1,0,0,0,0,0,1, ... 0,0,0,0,0,0,1,0,0,1,2,2,3,8,2
..      .      .      .      .      .      .      .      .      .      .

1,1,0,0,0,1,0,0,0,1,0,0,3,1,1,0,1,1,0,0,1,0,0,0,0,0,1,... 0,0,0,0,0,0,1,0,0,1,1,1,3,1,1
1,1,0,1,1,1,1,0,1,0,1,0,1,2,1,0,1,1,1,1,1,1,1,1,0,1, ... 1,0,1,0,0,0,1,0,0,0,1,1,2,7,1
    
```

Рисунок 2 – Фрагмент файла students.data
Figure 2 – A fragment of the students.data file

Каждому объекту (ответу обучаемого на вопросы анкеты) в файле данных соответствует отдельная строка. Так как наименование целевого признака (SATF) находится на первой строке файла students.names, то и значение этого признака расположено в начале каждой строки. Затем через запятую следуют значения всех остальных признаков.

Помимо рассмотренных в See5 в качестве дополнительных входных наборов данных могут быть использоваться файлы с расширениями .costs и .test.

Файл тестовых данных (расширение .test) используется для проверки качества построенного дерева решений. Формат этого файла совпадает с форматом файла данных (.data). В тестовом наборе данных содержатся новые объекты, например, в нашем случае это могут быть ответы обучаемых, полученные в результате опроса в следующем (предыдущем) по отношению к анализируемому учебном году.

Файл стоимости (расширение .costs) содержит «стоимость» различных ошибок классификации, например, в нашем случае можно установить стоимость ошибки при отнесении объекта к первому классу – «полностью удовлетворен», в то время как на самом деле это объект второго класса – «частично удовлетворен» выше чем стоимость ошибки при обратной ситуации (второй класс в результате применения дерева решений, а на самом деле – первый). Такое соотношение стоимостей ошибок соответствует характеру исследования, направленному на выявление причин условий неудовлетворенности обучаемых.

Интерфейс системы See5 подробно рассмотрен в [10]. В данной работе ограничимся только иллюстрацией выбора некоторых параметров (опций) построения дерева решений и получения результатов.

После загрузки данных в See5 и запуска режима «Construct Classifier» необходимо указать параметры построения дерева решений. Рассмотрим три варианта: 1) по умолчанию; 2) с отсеиванием переменных (опция «Winnow attributes»); 3) усиление решения (опция «Boost»).

```
See5 [Release 2.11a]    Fri Aug 23 09:20:56 2019
-----
Class specified by attribute `SATF'
Read 300 cases (218 attributes) from students 2018.data
Decision tree:
C.2_48 in {3,5}: 2 (13/1)
C.2_48 = 1:
...C.3_33->6 = 5: 1 (0)
:   C.3_33->6 in {1,3,4}: 1 (172/12)
:   C.3_33->6 = 2: 2 (51/22)
C.2_48 = 2:
...C.2_30->5 in {2,3,4,5}: 2 (48/4)
   C.2_30->5 = 1:
   ...C.3_32->5 = 0: 1 (10)
   C.3_32->5 = 1: 2 (6/1)
Evaluation on training data (300 cases):
Decision Tree
-----
Size      Errors
  6      40(13.3%)  <<
(a)      (b)      <-classified as
-----
  170     28      (a): class 1
  12      90      (b): class 2
Attribute usage:
  100%   C.2_48
   74%   C.3_33->6
   21%   C.2_30->5
   5%    C.3_32->5
Time: 0.1 secs
```

Рисунок 3 – Результаты See5 при настройках классификации по умолчанию
Figure 3 – See5 results with default classification settings

При первом варианте параметры «Construct Classifier Options» без изменений) система See5 выдаст следующие результаты (Рисунок 3).

Показателем точности полученного дерева решений являются ошибки классификации (Errors), которые в данном случае составляют 13,3 %, при 40 случаях ошибочной классификации.

Результат построения дерева решений с отсеиванием переменных следует (в окне «Construct Classifier Options» установлен параметр «Winnow attributes») указан на Рисунке 4. По сравнению с предыдущим (параметры классификации без изменений) ошибки классификации (Errors) уменьшились до 10,3 % при 31 случае ошибочной классификации. При этом в начале отчета (рис. 5) отмечено использование опции «Winnow attributes» и указано, что 199 признаков было отсеяно (199 attributes winnowed) и указана предполагаемая важность оставшихся признаков (Estimated importance of remaining attributes). Полученное с использованием данного параметра дерево решений несколько отличается от предыдущего и содержит больше «ветвей» для классификации по классу SATF = 1.

Основным предназначением отсеивания признаков является отделение полезных признаков, входящих в состав дерева решений, от бесполезных, не участвующих в классификации. Наряду со снижением размерности данный механизм можно применить и для выбора необходимого набора признаков, чтобы в дальнейшем уделить им больше

внимания в ходе формирования исходных наборов данных. Так, в нашем случае, отсеивание можно применить в дальнейшем для совершенствования и переработки вопросов анкеты путем исключения из опроса вопросов, не имеющих значения для целевого признака.

Третий вариант построения дерева решений основан на включении режима усиления решения (англ. boosting). Включение усиления осуществляется путем установки параметра «Boost» и указанием общего числа деревьев решений, получаемых в ходе See5. Результат полученный при включенном параметре «Boost» и значении «trials» равном 3 приведен на Рисунке 5.

Сравнивая последние результаты классификации с предыдущими (параметры по умолчанию и отсеивание признаков) можно отметить, что итоговые ошибки классификации (Errors, значение в строке boost) составляют 9,7 % при 29 случаях ошибочной классификации.

```
See5 [Release 2.11a]    Fri Aug 23 09:55:26 2019
-----
Options: winnow attributes
Class specified by attribute `SATF'
Read 300 cases (218 attributes) from students 2018.data
199 attributes winnowed
Estimated importance of remaining attributes:
 48% C.2_48
  7% C.2_30->5
  3% C.3_33->6
<1% C.1_28->7
<1% C.2_30->7
<1% C.3_32->5
<1% D_36->n/null
<1% F_49->4
<1% F_51->10
<1% G_54
Decision tree:
C.2_48 in {3,5}: 2 (13/1)
C.2_48 = 2:
...C.2_30->5 in {2,3,4,5}: 2 (48/4)
:   C.2_30->5 = 1:
:   ...C.3_32->5 = 0: 1 (10)
:   :   C.3_32->5 = 1: 2 (6/1)
C.2_48 = 1:
...C.3_33->6 = 5: 1 (0)
C.3_33->6 in {1,3,4}: 1 (172/12)
C.3_33->6 = 2:
...C.2_30->5 in {2,5}: 2 (29/8)
C.2_30->5 in {3,4}: 1 (7/2)
C.2_30->5 = 1:
...C.3_32->5 = 0: 2 (7/2)
:   C.3_32->5 = 1: 1 (8/1)
Evaluation on training data (300 cases):
Decision Tree
-----
Size      Errors
  9      31(10.3%)  <<
(a)      (b)      <-classified as
-----
 182     16      (a): class 1
  15     87      (b): class 2
Attribute usage:
 100%   C.2_48
  74%   C.3_33->6
  38%   C.2_30->5
  10%   C.3_32->5
Time: 0.1 secs
```

Рисунок 4 – Результаты See5 в случае выбора параметра «Winnow attributes»
Figure 4 – See5 results if the «Winnow attributes» parameter is selected


```

Class specified by attribute `SATF'
Read 300 cases (218 attributes) from students 2018.data
----- Trial 0: -----
Decision tree:
C.2_48 in {3,5}: 2 (13/1)
C.2_48 = 1:
:...C.3_33->6 = 5: 1 (0)
:  C.3_33->6 in {1,3,4}: 1 (172/12)
:  C.3_33->6 = 2: 2 (51/22)
C.2_48 = 2:
:...C.2_30->5 in {2,3,4,5}: 2 (48/4)
      C.2_30->5 = 1:
        :...C.3_32->5 = 0: 1 (10)
        :  C.3_32->5 = 1: 2 (6/1)
----- Trial 1: -----
Decision tree:
A_11 <= 1: 1 (112/8.7)
A_11 > 1:
:...A_11 > 3: 2 (9.5)
      A_11 <= 3:
        :...C.2_30->3 = 1: 1 (70.3/15.8)
        :  C.2_30->3 = 2: 2 (84.5/35.6)
        :  C.2_30->3 = 3: 2 (13.4)
        :  C.2_30->3 = 4: 1 (8.7/1.6)
        :  C.2_30->3 = 5: 2 (1.6)
----- Trial 2: -----
Decision tree:
C.2_48 = 2: 2 (73.7/14.4)
C.2_48 = 3: 2 (12.4)
C.2_48 = 5: 1 (3.3/1.3)
C.2_48 = 1:
:...C.2_30->7 = 5: 1 (0)
      C.2_30->7 = 1: 1 (126.5/10.5)
      C.2_30->7 = 2: 2 (57.8/24.3)
      C.2_30->7 = 3: 2 (3.3/0.7)
      C.2_30->7 = 4: 1 (7.9/2)
Evaluation on training data (300 cases):
Trial      Decision Tree
-----
          Size      Errors
0          6      40(13.3%)
1          7      48(16.0%)
2          7      58(19.3%)
boost      29( 9.7%)    <<
          (a)      (b)    <-classified as
          -----
          176     22      (a): class 1
          7       95      (b): class 2
Attribute usage:
100%  A_11
100%  C.2_48
74%   C.2_30->7
74%   C.3_33->6
56%   C.2_30->3
21%   C.2_30->5
5%    C.3_32->5
Time: 0.2 secs

```

Рисунок 5 – Результаты See5 при установленном «Boost» и «trials» = 3
Figure 5 – See5 results with «Boost» and «trials» = 3 set

Смысл данного режима заключается в построении не одного, как в первых двух случаях, а нескольких деревьев решений, которые как можно меньше дублируют друг друга. При этом пересечение множеств признаков, входящих в состав нескольких деревьев должны быть минимальным [10]. В нашем случае при указании в настройках классификатора различных значений параметра «trials» установлено, что наилучшие результаты с ошибкой 5,7 % и 17 случаями ошибочной классификации показывает число деревьев равное 5. Однако уменьшение доли ошибочных классификаций при увеличении числа деревьев приводит к существенному увеличению числа решающих правил и количества задействованных в ходе классификации признаков. При 5 деревьях число

правил (Size в отчете) составляет 50, а число используемых признаков равно 23 (Attribute usage в отчете), в то время как при 3 деревьях получено 20 правил при 7 используемых признаках. Увеличение числа деревьев значительно усложняет дальнейшую интерпретацию и применение результатов.

Изменение других настроек алгоритма классификации (параметры в окне «Construct Classifier Options») является предметом дальнейшего, более глубокого анализа, проведение которого целесообразно после корректировки структуры анкеты с учетом результатов, полученных в ходе данного исследования, и сбора необходимого объема данных об удовлетворенности обучаемых в ходе очередного опроса в текущем учебном году.

Обсуждение

Для представления полученных результатов анализа удовлетворенности обучением рассмотрим последний вариант построения классификационных правил с использованием 3 деревьев решений (режим boosting) для ответа на вопросы о том, какие элементы и условия реализации образовательных программ оказывают существенное влияние на оценку удовлетворенности обучаемых получаемым образованием, какие значения отмеченных признаков обеспечивают отнесение объекта к одному из классов. Полученные результаты имеют меньшую ошибку 9,7 % против 10,3 % и 13,3 % в первых двух вариантах конструирования классификатора. А использование 7 признаков против 4 учитывает больше признаков, характеризующих деятельность образовательной организации.

Отчет системы See5 для рассматриваемого варианта (Рисунок 5) включает блок используемых признаков (Attribute usage). В данном блоке приведены наименования признаков в соответствии с файлом .names (правая колонка) и доля их участия в классификации (левая колонка). Так в соответствии со структурой, используемой в нашем случае анкеты, к указанным классификационным признакам относятся:

- 1) 100 % A_11 – престиж обучения в институте (ответ на вопрос «Как Вы считаете, насколько сегодня престижным является обучение в Институте?»);
- 2) 100 % C.2_48 – характер атмосферы межличностных отношений в институте (ответ на вопрос «Можно ли утверждать, что в Институте в целом создана и поддерживается благоприятная для обучения атмосфера межличностных отношений?»);
- 3) 74 % C.2_30->7 – оценка качества проводимых учебных занятий (ответ на вопрос «Насколько Вы удовлетворены качеством проводимых учебных занятий?»);
- 4) 74 % C.3_33->6 – оценка доступности информации о порядке прохождения промежуточной аттестации (ответ на вопрос «Оцените, насколько Вы удовлетворены доступностью информации о порядке прохождения промежуточной аттестации (экзамены, зачеты)?»);
- 5) 56 % C.2_30->3 – признание достижений обучаемого (ответ на вопрос «Насколько Вы удовлетворены признанием Ваших успехов и достижений?»);
- 6) 21 % C.2_30->5 – оценка деятельности руководства института (ответ на вопрос «Насколько Вы удовлетворены деятельностью руководства Института?»);
- 7) 5 % C.3_32->5 – характер информационной среды института (выбор ответа «Информационные стенды и объявления» на вопрос «Укажите, пожалуйста, из каких источников Вы получаете информацию о жизни нашего Института?»).

Другой вопрос, ответ на который можно получить на основе отчета See5, - каковы значения классификационных признаков (характера ответов, полученных в ходе опроса), обеспечивающих отнесение объекта к тому или иному классу. Так, в нашем случае, интерес представляют значения квалификационных признаков для класса 1 «Полностью

удовлетворен». Для определения таких значений воспользуемся ветками деревьев решений для целевого признака $SATF = 1$ (Рисунок 5). Запишем эти структуры в виде более понятных логических правил (Таблица 1) и пояснений по соответствующим значениям классификационных признаков.

Таблица 1 – Логические правила классификации и пояснения значений признаков для класса «Полностью удовлетворен»

Table 1 – Logical rules for the classification and explanation of the values of the characteristics for the class «fully satisfied»

Логическое правило	Пояснение значений признаков
ЕСЛИ $C.2_{48} = 1$ И $C.3_{33} \rightarrow 6 = 1$ ТО $SATF = 1$	На вопрос «Можно ли утверждать, что в Институте в целом создана и поддерживается благоприятная для обучения атмосфера межличностных отношений?» дается ответ «Да, вполне» и на вопрос «Оцените, насколько Вы удовлетворены доступностью информации о порядке прохождения промежуточной аттестации (экзамены, зачеты)?» даются ответ «Полностью удовлетворен».
ЕСЛИ $C.2_{48} = 2$ И $C.3_{30} \rightarrow 5 = 1$ И $C.3_{32} \rightarrow 5 = 0$ ТО $SATF = 1$	На вопрос «Можно ли утверждать, что в Институте в целом создана и поддерживается благоприятная для обучения атмосфера межличностных отношений?» дается ответ «В целом такая атмосфера создана, но временами в психологическом плане бывает тяжело» и на вопрос «Насколько Вы удовлетворены деятельностью руководства Института?» дается ответ «Полностью удовлетворен» и на вопрос «Укажите, пожалуйста, из каких источников Вы получаете информацию о жизни нашего Института?» ответ «Информационные стенды, объявления» не выбирается (т.е. как источник информации не используется).
ЕСЛИ $A_{11} \leq 1$ ТО $SATF = 1$	На вопрос «Как Вы считаете, насколько сегодня престижным является обучение в Институте?» дается ответ «Очень престижно»
ЕСЛИ $A_{11} > 1$ И $A_{11} \leq 3$ И $C.2_{30} \rightarrow 3 = 1$ ТО $SATF = 1$	На вопрос «Как Вы считаете, насколько сегодня престижным является обучение в Институте?» даются ответы «Скорее престижно, чем не престижно», «Скорее не престижно, чем престижно» и на вопрос «Насколько Вы удовлетворены признанием Ваших успехов и достижений?» дается ответ «Полностью удовлетворен».
ЕСЛИ $C.2_{48} = 1$ И $C.2_{30} \rightarrow 7 = 1$ ТО $SATF = 1$	На вопрос «Можно ли утверждать, что в Институте в целом создана и поддерживается благоприятная для обучения атмосфера межличностных отношений?» дается ответ «Да, вполне» и на вопрос «Насколько Вы удовлетворены качеством проводимых учебных занятий?» дается ответ «Полностью удовлетворен».

Аналогично представим логические правила для класса 2 «Частично удовлетворен» – $SATF = 2$ (Таблица 2).

Таблица 2 – Логические правила классификации и пояснения значений признаков для класса «Частично удовлетворен»

Table 2 – Logical rules for the classification and explanation of the values of characteristics for the class «Partially satisfied»

Логическое правило	Пояснение значений признаков
ЕСЛИ $C.2_{48} = 3$ ИЛИ $C.2_{48} = 5$ ТО $SATF = 2$	На вопрос «Можно ли утверждать, что в Институте в целом создана и поддерживается благоприятная для обучения атмосфера межличностных отношений?» даются ответы «Нет, атмосфера скорее неблагоприятная, чем благоприятная» или «Нет, атмосфера межличностных отношений в институте совершенно неблагоприятна для процесса обучения».
ЕСЛИ $C.2_{48} = 1$ И $C.3_{33} > 6 = 2$ ТО $SATF = 2$	На вопрос «Можно ли утверждать, что в Институте в целом создана и поддерживается благоприятная для обучения атмосфера межличностных отношений?» дается ответ «Да, вполне» и на вопрос «Оцените, насколько Вы удовлетворены доступностью информации о порядке прохождения промежуточной аттестации (экзамены, зачеты)?» даются ответ «Скорее удовлетворен».
ЕСЛИ $C.2_{48} = 2$ И $C.3_{30} > 5 < > 1$ ТО $SATF = 2$	На вопрос «Можно ли утверждать, что в Институте в целом создана и поддерживается благоприятная для обучения атмосфера межличностных отношений?» дается ответ «В целом такая атмосфера создана, но временами в психологическом плане бывает тяжело» и на вопрос «Насколько Вы удовлетворены деятельностью руководства Института?» любые ответы кроме «Полностью удовлетворен».
ЕСЛИ $A_{11} > 3$ ТО $SATF = 2$	На вопрос «Как Вы считаете, насколько сегодня престижным является обучение в Институте?» даются ответы «Совсем не престижно» и «Затрудняюсь ответить»
ЕСЛИ $A_{11} > 1$ И $A_{11} \leq 3$ И ($C.2_{30} > 3 = 2$ ИЛИ $C.2_{30} > 3 = 3$) ТО $SATF = 2$	На вопрос «Как Вы считаете, насколько сегодня престижным является обучение в Институте?» даются ответы «Скорее престижно, чем не престижно», «Скорее не престижно, чем престижно» и на вопрос «Насколько Вы удовлетворены признанием Ваших успехов и достижений?» дается ответ «Скорее удовлетворен» или «Скорее не удовлетворен».
ЕСЛИ $C.2_{48} = 2$ ИЛИ $C.2_{48} = 3$) ТО $SATF = 2$	На вопрос «Можно ли утверждать, что в Институте в целом создана и поддерживается благоприятная для обучения атмосфера межличностных отношений?» дается ответ «В целом такая атмосфера создана, но временами в психологическом плане бывает тяжело» или «Нет, атмосфера скорее неблагоприятная, чем благоприятная».
ЕСЛИ $C.2_{48} = 1$ И ($C.2_{30} > 7 = 2$ ИЛИ $C.2_{30} > 7 = 3$) ТО $SATF = 1$	На вопрос «Можно ли утверждать, что в Институте в целом создана и поддерживается благоприятная для обучения атмосфера межличностных отношений?» дается ответ «Да, вполне» и на вопрос «Насколько Вы удовлетворены качеством проводимых учебных занятий?» дается ответ «Скорее удовлетворен» или «Скорее не удовлетворен».

С учетом приведенных в Таблице 1 значений классификационных признаков, обеспечивающих отнесение объекта к классу 1 «Полностью удовлетворен», можно отметить, что высокая оценка удовлетворенности обучением в нашем случае определяется прежде всего высоким престижем обучения в институте, благоприятным характером атмосферы межличностных отношений, высоким уровнем качества проводимых учебных занятий, доступностью информации о порядке прохождения промежуточной аттестации, признанием достижений обучаемого, а также высокой оценкой деятельности руководства института.

В тоже время, приведенные в Таблице 2 значения квалификационных признаков, обеспечивающих отнесение объекта к классу 2 «Частично удовлетворен», показывают, что снижение оценки удовлетворенности обучением в нашем случае связано с неблагоприятной атмосферой межличностного общения, неполнотой (низким качеством) информации о порядке прохождения промежуточной аттестации, недостаточным признанием личных успехов и достижений обучаемого, неполная удовлетворенность учебными занятиями и деятельностью руководства института.

Интересной возможностью See5 с точки зрения применения полученных классификаторов является предсказание удовлетворенности обучаемых получаемым образованием на основе предположений об изменениях, проводимых в институте по различным направлениям. Данные предположения могут быть представлены в форме ответов на вопросы анкеты и выражены определенными значениями классификационных признаков. В режиме использования классификатора (Use Classifier) See5 предлагает последовательно ввести соответствующие значения классификационных признаков, а затем выводит предсказание класса – значение целевого признака.

Например, предположим следующие значения квалификационных признаков и соответствующие им ответы на вопросы анкеты (Таблица 3)

Таблица 3 – Предполагаемые значения квалификационных признаков
Table 3 – Estimated values of qualification characteristics

Наименование признака	Ответ на вопрос анкеты с учетом предположений о состоянии образовательной деятельности института
С.2_48 = 2	На вопрос «Можно ли утверждать, что в Институте в целом создана и поддерживается благоприятная для обучения атмосфера межличностных отношений?» дается ответ «В целом такая атмосфера создана, но временами в психологическом плане бывает тяжело». Такой ответ обучаемых означает, что текущее состояние межличностных отношений включает элементы принуждения к выполнению ряда обязательных задач, связанных спецификой института (служба в нарядах, участие в различных мероприятиях).
С.2_30->5 = 1	на вопрос «Насколько Вы удовлетворены деятельностью руководства Института?» дается ответ «Полностью удовлетворен». При этом обучаемые видят и положительно оценивают ключевую роль руководства в организации и контроле всех направлений деятельности института.
С.3_32->5 = 0	на вопрос «Укажите, пожалуйста, из каких источников Вы получаете информацию о жизни нашего Института?» ответ «Информационные стенды, объявления» не выбирается. Такая позиция в ответе обучаемых реализуется за счет отказа от размещения информации на

	различных стендах и в объявлениях, а информирование осуществляется в основном посредством электронных каналов.
A_11 = 3	На вопрос «Как Вы считаете, насколько сегодня престижным является обучение в Институте?» дается ответ «Скорее не престижно, чем престижно». Выбор этого варианта ответа может быть связан со снижением престижа службы в полиции в целом (высокая нагрузка на сотрудников, невысокий уровень денежного довольствия, снижение доверия общества).
C.2_30->3 = 2	на вопрос «Насколько Вы удовлетворены признанием Ваших успехов и достижений?» дается ответ «Скорее удовлетворен». Данное значение является результатом изменения политики оценивания деятельности обучаемых в сторону учета не только показателей успеваемости, но и участия в общественной жизни, спорте, науке, личной ответственности.

С учетом значений, указанных в Таблице 3, See5 выдаст полученное значение целевого признака – SATF = 2. Это значит, что в данной ситуации оценка удовлетворенности обучаемых получаемым образованием относится к классу 2 «Частично удовлетворен». При этом See5 также указывает вероятность отнесения к классу 2, которая составляет 0,62.

Заключение

Таким образом, рассмотренный в работе подход к анализу результатов опроса обучаемых на основе применения инструментария интеллектуального анализа – приложения See5 – позволяет за короткое время обработать большой объем данных и получить ряд интересных результатов. К таким результатам можно, прежде всего, отнести определение условий и направлений, оказывающих влияние на оценку удовлетворенности обучаемых от получаемого образования, и их учет в ходе принятия решений по совершенствованию образовательной деятельности института. Возможности предсказания класса целевого признака в зависимости от предположений о значениях квалификационных признаков (ответов обучаемых на вопросы анкеты) обеспечивают возможности выдвижения гипотез при оценке предстоящих изменений в характере реализации направлений деятельности института. Дальнейшая работа в данной области, по мнению авторов, подразумевает с одной стороны совершенствование структуры анкеты в части более четких формулировок вопросов по полученным условиям и направлениям, а также с учетом результатов отсеивания классификационных признаков. С другой стороны, следует уделить внимание разработке обобщенного критерия оценки удовлетворенности обучаемых. При этом необходимо в качестве основы использовать представленную в работе группировку вопросов анкеты по направлениям образовательной деятельности института.

ЛИТЕРАТУРА

1. Клейменов Е.А. Институциональные ожидания курсантов и слушателей в образовательной организации. *Власть и управление на Востоке России*. 2018;3:92-105.
2. Русанова А.А. Удовлетворенность как показатель качества образования в современном вузе (из опыта конкретного социологического исследования).

- Современные исследования социальных проблем (электронный научный журнал)*. 2011;3:21-26.
3. Киселёва М.В. Вовлечение обучающихся вуза в оценку и повышение качества образования. *Вестник Костромского государственного университета. Серия: Педагогика. Психология. Социокинетика*. 2016;22(3):142-147.
 4. Бондарь К.М., Скрипко П.Б. Анализ данных в образовании как основа для повышения качества образовательного контента и эффективности обучения. *Социально-психологические и педагогические детерминанты формирования и развития профессиональной мотивации курсантов и слушателей в процессе обучения в образовательных организациях системы МВД России* Материалы всероссийской научно-практической конференции. Отв. ред. А.А. Андреев. 2017:130-133.
 5. Pappas Ch. 9 Tips To Give and Receive eLearning Feedback. Доступно по: <https://elearningindustry.com/9-tips-give-receive-elearning-feedback>. Ссылка активна на 02.09.2019.
 6. Скрипко П.Б., Дунин В.С. Направления применения методов интеллектуального анализа данных для оценки результатов электронного обучения. *Государство и право в эпоху глобальных перемен материалы международной научно-практической конференции*. Под ред. С.К. Бурякова. 2017:307-308.
 7. Agaoglu M. Predicting Instructor Performance Using Data Mining. *IEEE Access*. 2016;4:2379-2387.
 8. Дюк В., Самойленко А. *Data Mining: учебный курс*. СПб.: Питер; 2001:366.
 9. Барсегян А.А. и др. *Анализ данных и процессов: учеб. пособие*. СПб.: БХВ-Петербург; 2009:512.
 10. Бондарь К.М., Емельянова О.В., Скрипко П.Б. *Методика проведения аналитической работы в ОВД: учебно-практическое пособие*. Дальневосточный юрид. ин-т МВД РФ. Хабаровск: РИО ДВЮИ МВД РФ; 2013:144.

REFERENCES

1. Kleimenov E.A. Institutional expectations of cadets and students in an educational organization. *Power and Management in the East of Russia*. 2018;3:92-105. (In Russ.).
2. Rusanova A.A. Satisfaction as an indicator of the quality of education in a modern university (from the experience of a specific sociological study). *Modern Research on Social Problems (electronic scientific journal)*. 2011;3:21-26. (In Russ.).
3. Kiseleva M.V. Involving university students in assessing and improving the quality of education. *The Bulletin of Kostroma State University. Series: Pedagogy. Psychology. Sociokinetics*. 2016;22(3):142-147. (In Russ.).
4. Bondar K.M., Skripko P.B. Analysis of data in education as a basis for improving the quality of educational content and the effectiveness of training. *Socio-psychological and pedagogical determinants of the formation and development of professional motivation of cadets and students in the learning process in educational organizations of the Ministry of Internal Affairs of Russia* Materials of the All-Russian Scientific and Practical Conference. Executive Editor A.A. Andreev. 2017:130-133. (In Russ.).
5. Pappas Ch. 9 Tips To Give and Receive eLearning Feedback. Available by: <https://elearningindustry.com/9-tips-give-receive-elearning-feedback> (accessed September 2, 2019).
6. Skripko P.B., Dunin V.S. Directions for the application of data mining methods for evaluating the results of e-learning. *In the collection: State and law in the era of global change, materials of the international scientific and practical conference*. Edited by S.K. Buryakova. 2017:307-308. (In Russ.).

7. Agaoglu M. Predicting Instructor Performance Using Data Mining. *IEEE Access*. 2016;4:2379-2387.
8. Duke V., Samoilenko A. Data Mining: training course. St. Petersburg: Peter;2001:366. (In Russ.).
9. Barseghyan A.A. et al. Analysis of data and processes: textbook allowance. St. Petersburg: BHV-Petersburg;2009:512. (In Russ.).
10. Bondar K.M., Emelyanova O.V., Skripko P.B. Methodology for conducting analytical work in the Department of Internal Affairs: a training manual. Far Eastern legal Institute of the Ministry of Internal Affairs of the Russian Federation. Khabarovsk: RIO DVUI of the Ministry of Internal Affairs of the Russian Federation;2013:144. (In Russ.).

ИНФОРМАЦИЯ ОБ АВТОРЕ / INFORMATION ABOUT THE AUTHOR

Скрипко Павел Борисович, канд. техн. наук, доцент, начальник кафедры информационного и технического обеспечения ОВД, ФГКОУ ВО Дальневосточный юридический институт МВД России, Хабаровск, Российская Федерация.

Дунин Вадим Сергеевич, канд. техн. наук, заместитель начальника кафедры информационного и технического обеспечения ОВД, ФГКОУ ВО Дальневосточный юридический институт МВД России, Хабаровск, Российская Федерация.

Клеймёнов Евгений Алексеевич, канд. соц. наук, доцент кафедры социально-гуманитарных и экономических дисциплин, ФГКОУ ВО Дальневосточный юридический институт МВД России, Хабаровск, Российская Федерация.

Pavel B. Skripko, Candidate of Engineering Sciences, Associate Professor, Head of the Department of Information and Technical Support of the Internal Affairs Bodies, Far Eastern Law Institute of the Ministry of the Interior of the Russian Federation, Khabarovsk, the Russian Federation.

Vadim S. Dunin, Candidate of Engineering Sciences, Deputy Head of the Department of Information and Technical Support of Internal Affairs Bodies, Far Eastern Law Institute of the Ministry of the Interior of the Russian Federation, Khabarovsk, the Russian Federation.

Eugene A. Kleimenov, Candidate of Sociological Sciences, Associate Professor, Department of Social and Humanitarian and Economic Disciplines, Far Eastern Law Institute of the Ministry of the Interior of the Russian Federation, Khabarovsk, the Russian Federation.