МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М.В.Ломоносова»

«Утверждаю»

Декан факультета ВМК МГУ
имени М.В. Ломоносова

академик _____ Е.И.Моисеев

«___» _____ 2017 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

«Технологии прикладного анализа данных SAS»

Уровень высшего образования – подготовка научно-педагогических кадров в аспирантуре

Направление подготовки – 09.06.01 «Информатика и вычислительная техника»

Направленность (профиль) – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей» (05.13.11)

2017 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

1. Наименование дисциплины
ТЕХНОЛОГИИ ПРИКЛАДНОГО АНАЛИЗА ДАННЫХ SAS

2. Уровень высшего образования
Подготовка научно-педагогических кадров в аспирантуре.

3. Направление подготовки, направленность (профиль) подготовки
Направление подготовки: 09.06.01 «Информатика и вычислительная техника»;
Направленность (профиль):«Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей» (05.13.11).

4. Место дисциплины в структуре основной образовательной программы
Дисциплина относится к специальным курсам по выбору вариативной части образовательной программы.

5. Перечень планируемых результатов обучения
Дисциплина участвует в формировании следующих компетенций образовательной программы:

Формируемые компетенции *(код компетенции)*	**Планируемые результаты обучения по дисциплине (модулю)**
Владение методологией теоретических и экспериментальных исследований в области профессиональной деятельности (ОПК-1)	ЗНАТЬ: классические математические методы, применяющиеся для решения задач в области естественных наук, экономики, социологии и информационно-коммуникационных технологий; УМЕТЬ: применять классические методы построения и анализа математических моделей; ВЛАДЕТЬ: базовыми навыками выбора методов и средств построения и анализа математических моделей

Владение современными методами построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также методами разработки и реализации алгоритмов их решения на основе фундаментальных знаний в области математики и информатики (ПК-1)	ЗНАТЬ: классические методы построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также базовые методы разработки и реализации алгоритмов их решения; УМЕТЬ: применять классические методы построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также базовые методы разработки и реализации алгоритмов их решения; ВЛАДЕТЬ: базовыми навыками выбора методов построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также методов разработки и реализации алгоритмов их решения.
Владение современными алгоритмами разработки программного обеспечения вычислительных комплексов (ПК-3)	ЗНАТЬ: современные алгоритмы разработки программного обеспечения вычислительных комплексов; УМЕТЬ: применять современные алгоритмы разработки программного обеспечения вычислительных комплексов; ВЛАДЕТЬ: базовыми навыками выбора современных алгоритмов разработки программного обеспечения вычислительных комплексов.
Владение современными методами интеллектуального анализа данных (ПК-5)	ЗНАТЬ: современные методы интеллектуального анализа данных; УМЕТЬ: применять современные методы интеллектуального анализа данных; ВЛАДЕТЬ: базовыми навыками выбора методов интеллектуального анализа данных.

Оценочные средства для промежуточной аттестации приведены в Приложении.

6. Объем дисциплины

Объем дисциплины составляет **3**зачетных единицы, всего **108** часов.

72часа составляет контактная работа с преподавателем (**40**часов занятий лекционного типа, **30**часов занятий семинарского типа (семинары, научно-практические занятия, лабораторные работы и т.п.), **0** часов групповых консультаций, **0** часовиндивидуальных консультаций, **0**часов мероприятий текущего контроля успеваемости, **2** часа мероприятий промежуточной аттестации).

36 часов составляетсамостоятельная работа учащихся.

7. Входные требования для освоения дисциплины

Учащиеся должны владеть знаниями по базам данным и языкам программирования, а также по математической статистике в объеме, соответствующем основным образовательным программам бакалавриата и магистратуры по укрупненным группам направлений и специальностей 01.00.00 «Математика и механика», 02.00.00 «Компьютерные и информационные науки».

8. Образовательные технологии

В процессе обучения используется бесплатная аналитическая платформа SASUniversityEdition, включающая средства программирования для обработки данных разной структуры и сложности, а также библиотеку методов для статистического анализа данных. Учебный курс состоит из двух основных блоков. Первый блок посвящен обучению программированию на языках аналитической платформы SAS (Base, Macro, SQL), второй блок– изучению возможностей библиотеки статистического анализа данных SASSTATдля решения типовых задач регрессионного, дисперсионного анализа данных и проверки статистических гипотез. В рамках курса читаются лекции, и проводятся практические занятия, включая выполнение самостоятельных практических заданий по разработке моделей для статистического анализа данных. Также в качестве вспомогательных форм обучения аспиранты могут пользоваться бесплатными онлайн курсами на английском языке:

- SAS Programming I: Essentials https://support.sas.com/edu/schedules.html?ctry=us&id=277;
- SAS Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression https://support.sas.com/edu/schedules.html?ctry=us&id=1979.

9. Содержание дисциплины

В курсе рассматриваются основные вопросы программирования для решения задач статистического анализа данных с использованием аналитической платформы SAS. В первой части курса, посвященной обучению программированию для решения задач подготовки данных и

формирования отчетности, рассматриваются: основные принципы работы шага обработки данных; работа со структурированными наборами данных и массивами; форматы и типы данныхязыка SASBase; процедуры преобразования форматов и типов; работа с внешними сложно структурированными наборами данных; алгоритмы и методы для организации поиска по ключу с помощью индексов, форматов, хэш-объектов; методы разработки и использования пользовательских процедур и функций; программирование с использованием макросов,макропеременных и макроподстановок; использованием языка SQL; формирование отчетов и работа подсистемой вывода; графические возможности и процедуры. Вторая часть курса посвящена изучению методов разработки программ для статистического анализа данных с использованием соответствующих библиотек аналитической платформы SAS. Рассматриваются следующие вопросы: процедуры и методы для проверки статистических гипотез; модели и процедуры для дисперсионного анализа данных; построениелинейных регрессионных моделей;проблема мультиколлинеарности; методы пошагового отбора переменных, регуляризации, преобразования пространства признаков; процедуры поиска главных компонент и кластеризации переменных;процедуры и инструменты для поиска выбросов; процедуры построения нелинейных регрессий; анализ таблиц сопряженности; логистическая регрессия; обобщенные линейные модели, пуассоновская и гамма регрессии; методы сравнения и оценки моделей на тестовом наборе данных.Демонстрация примеров использования изучаемых методов и процедур проводится преподавателями на каждой лекции и каждом семинаре. Также данная дисциплина поддерживается практическими заданиями (практическими самостоятельными работами), позволяющими аспирантам овладеть навыками написания программ для статистической обработки данных, а также навыками анализа результатов работы реализованных алгоритмов. Обсуждение практических самостоятельных работ, а также их защита, проводятся на семинарах. Дополнительно, на семинарах аспиранты выполняют небольшие практические задания по тематике последней на момент данного семинара лекции.Темы семинаров соответствуют темам лекций. Семинары направлены на укрепление знаний, полученных на лекциях.

Наименование и краткое содержание разделов и тем дисциплины (модуля), **форма промежуточной аттестации по дисциплине (модулю)**	**Всего (часы)**	В том числе								
		Контактная работа (работа во взаимодействии с преподавателем), часы из них						**Самостоятельная работа обучающегося, часы** из них		
		Занятия лекционного типа	Занятия семинарского типа	Групповые консультации	Индивидуальные консультации	Учебные занятия, направленные на проведение текущего контроля успеваемости коллоквиумы, практические контрольные занятия и др)*	**Всего**	Выполнение домашних заданий	Подготовка рефератови т.п..	**Всего**
Раздел1. Программирование на языках SASBase, Macro, SQL. Тема 1.1. Основные принципы работы шага обработки данных. Тема 1.2. Работа со структурированными наборами данных и массивами. Тема 1.3. Форматы и типы данных, процедуры преобразования форматов и типов.	45	20	15	–	–	–	35	10	0	10

Тема 1.4. Работа с внешними сложно структурированными наборами данных. Тема 1.5.Организация поиска по ключу с помощью индексов, форматов, хэш-объектов. Тема 1.6. Пользовательские процедуры и функции. Тема 1.7. Макропеременные и макроподстановки. Взаимодействие с SAS SQL. Тема 1.8. Работа с подсистемой вывода. Тема 1.9. Графические возможности и процедуры аналитической платформы.										
Тема 2.Основы прикладного статистического анализа данных на аналитической платформе SAS Тема 2.1. Процедуры проверки гипотез и дисперсионного анализа. Тема 2.2. Процедуры построения линейных регрессионных моделей. Смешанные линейные регрессионные модели. Тема 2.3. Проблема	45	20	15	–	–	–	35	10	0	10

мультиколлинеарности, пошаговый отбор переменных, регуляризация, преобразования пространства признаков. Тема 2.4. Процедуры поиска главных компонент и кластеризации переменных. Тема 2.5. Процедуры и инструменты для поиска выбросов. Тема 2.6. Процедуры построения нелинейных регрессий. Тема 2.7. Анализ таблиц сопряженности, логистическая регрессия. Тема 2.8. Обобщенные линейные модели, пуассоновская и гамма регрессии. Тема 2.9. Сравнение и оценка моделей на тестовом наборе данных. Тема 2.10. Методы кластеризации.										
Промежуточная аттестация– ЭКЗАМЕН	18	2						16		
Итого	108	72						36		

10. Учебно-методические материалы для самостоятельной работы учащихся

Самостоятельная работа учащихся проводится в виде выполнения практических самостоятельных работ (ПСР).

Текущий контроль осуществляется путем проверки ПСР преподавателями, а также индивидуального обсуждения (защиты) с преподавателями выполненной ПСР на семинарах. За каждую ПСР аспиранту проставляется определенное количество баллов. Итоговая сумма баллов влияет на итоговую оценку учащегосяпо данной дисциплине.
Также на семинарахаспиранты выполняют небольшие практические задания по тематике последней на момент данного семинара лекции. Качество выполнения заданий также влияет на итоговую оценку.

Методика выставления оценки по данной дисциплине (с учетом ПСР, а также практических заданий на семинарах) приведена в разделе«Методические материалы для проведения процедур оценивания результатов обучения» Приложения.

Типовые задания для ПСР и методические рекомендации к их выполнению

Домашнее задание №1 (ПСР №1)

Целью домашнего задания №1 является освоение языка SASBase для чтения структурированных наборов данных, разработки и применения форматов и пользовательских функций, методов расчета агрегационных показателей и формирование внешних сложно структурированных наборов данных.

Формулировка задания:

На основе набора данных SASHELP.CARS сформировать иерархический текстовый файл вида:

Производитель далее в скобках (Географическая область происхождения Производителя)

Число моделей Производителя по каждому числу цилиндров (CN) в строку вида: C4=X C12=y …

Средняя мощность для второго варианта по каждому числу цилиндров в строку.

При этом размер счета сохранять в рублях по курсу 57 рублей за доллар и писать в виде CN=XXXXXXX.XX RUB, а мощность сохранять в

ваттах в виде CN=XXXXX,XX Watt. Для этого реализовать необходимые форматы и FCMP функции для пересчета единиц измерения.

Домашнее задание №2 (ПСР №2)

Целью домашнего задания №2 является освоение работы на языках SASBaseи SASMacroс внешним сложно структурированным набором данных, реализация методов предобработки (в частности расчетов агрегатов, фильтрации и транспонирования данных) данных.

Формулировка задания:

Считать сформированный в первом задании текстовый файл и по нему сформировать набор данных с суммарной статистикой по среднему размеру счета «Число цилиндров»X«География» вида:

	Asia	USA	Europe	…
C4	X	Y	Z	…
C12	…	…	…	…

При этом единицы измерения должны использоваться исходные (доллары и лошадиные силы), для чтения применять шаблоны, для перевода единиц функции FCMP.

Домашнее задание №3 (ПСР №3)

Целью домашнего задания №3 является освоение процедур и методов проверки статистических гипотез и дисперсионного анализа аналитической платформы SAS, а также изучение процедур формирования отчетности и работы с подсистемой вывода.

Формулировка задания:

На основе набора данных SASHELP.CARS сделать следующее (Уровень значимости для всех гипотез задается 0.01):

1) Проверить влияет ли тип Кузова (Type) на расход бензина в городе (MPG_city) с заданным уровнем значимости.

2) Преобразовать категориальные переменные так, чтобы не было «неразличимых» групп (используя график diffogram, попарный t-test и шаг данных для объединения «неразличимых» групп). Написать соответствующий код.

3) Добавить предиктор Страна производитель, чтобы понять улучшается ли модель (по RMSE). Объединения «неразличимых» групп не проводить.

4) Проверить, нужен ли предиктор Origin*Type и если не нужен, то исключить из модели. Построить финальную модель.

5) Сформировать pdf отчет, содержащий ТОЛЬКО таблицы: «LeastSquaresMeansforeffect …» шага 2, ANOVA таблицу и таблицу статистик (fitstatistics) последнего шага.

Домашнее задание №4 (ПСР №4)

Целью домашнего задания №4 является освоение методов формирования шаблонов для гибкой настройки процедур работы с подсистемой вывода.

Формулировка задания:

Бонус (необязательное) задание (позволит сгладить недочеты в других частях задачи или в других задачах): Перед вторым шагом задания 3 написать шаблон с помощью процедуры template, чтобы в таблице попарных сравнений «LeastSquaresMeansforeffect …» (имя объекта stat.GLM.Pdiff) значение P-values больше заданного уровня значимости (0.01) подсвечивались красным или отмечалось «*».

Домашнее задание №5(ПСР №5)

Целью домашнего задания №5 является освоение методов построения полноценных регрессионных моделей, включая процедуры преобразования и предобработки данных, поиска и удаления выбросов, формирования, оценки и выбора моделей разной сложности

(включая обобщенные линейные модели), визуализации результатов.

Формулировка задания:

Вместе с заданием передается два макроса:

- prepare_data генерирует из набора данных cars случайным образом тренировочный и тестовый наборы (с сохранением распределения откликов). С тренировочным набором можно производить любые манипуляции (менять любые переменные, удалять записи и т.д.), в тестовом можно только дописывать входные переменные, например, X*X или f(X), причем по тем же правилам, что и в тренировочном наборе, менять переменные отклика в тестовом наборе нельзя.

- calc_mape считает на тестовом наборе оценку качества модели MAPE.

Задача состоит в том, чтобы построить регрессионную модель для прогнозирования расхода бензина на трассе (MPG_highway) от числовых переменных LengthWeightWheelbaseHorsepowerInvoiceEngineSizeCylinders и категориальных переменных Origin и Type. Можно использовать процедуры, рассмотренные в рамках курса: GLM, GLMSELECT, REG, GLMMOD, а также их комбинации, например, можно фильтровать переменные одной процедурой, а модель строить другой по отобранным переменным, или оценивать выбросы одной, потом их удалять из тренировочного набора, а строить модель другой процедурой по отфильтрованному набору. Для оценки качества разработанной модели использовать макрос calc_mape. Для корректной работы макроса необходимо, чтобы результат прогноза писался в переменную Result. Обратите внимание, что при разных запусках prepare_data генерируются разные тестовые и тренировочные наборы. Необходимо получать MAPE строго меньше 0.065. Рекомендуется использовать техники:

- Преобразования категориальных переменных (группировка значений) с использованием дисперсионного анализа (не забывайте делать эти преобразования и в тестовом наборе)

- Отбор значимых переменных с помощью REG или GLMSELECT (тестовый набор в качестве валидационного использовать нельзя, но можно из тренировочного выделить часть для валидации)

- Преобразование входных переменных и добавление полиномиальных членов в уравнение регрессии (не забывайте делать эти преобразования и в тестовом наборе)

- Преобразование отклика или использование обобщенных линейных моделей с разными распределениями ошибки и функциями связи

(не забывайте делать правильный пересчет отклика после прогноза).

Для получения требуемого качества модели обычно достаточно правильно использовать 1-2 из перечисленных выше техник.

Построить 3D график зависимости прогноза отклика от переменных Weight и Horsepower с равномерной сеткой 20 на 20 точек (сетку в наборе данных сгенерировать самостоятельно).

Данные ПСР соответствуют изучаемым темам следующим образом:

№	Изучаемая тема	Соответствующая ПСР
1	Темы1.1-1.4	ПСР №1
2	Темы 1.5-1.7	ПСР №2
3	Тема 2.1	ПСР №3
4	Тема 1.8 , Тема 2.1	ПСР №4
5	Тема 2.2-2.9, Тема 1.9	ПСР №5

11. Ресурсное обеспечение

Основная литература

1. С.А. Айвазян и др. /Прикладная статистика: Основы моделирования и первичная обработка данных. Справочное издание. - М.: Финансы и статистика, 1983. — 471 с.
2. С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. /Прикладная статистика: Исследование зависимостей.Справочное издание. - М.: Финансы и статистика, 1985. - 487с.
3. Айвазян С.А., Бухтштабер В.М., Енюков И.С., Мешалкин Л.Д. /Прикладная статистика: Классификации и снижение размерности.Справочное издание. - М.: Финансы и статистика, 1989. - 607с.
4. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. – Springer, 2001.
5. Lora D. Delwiche, Susan J. Slaughter The Little SAS® Book: A Primer, Fifth Edition - SAS Institute, 2012
6. Ron Cody An Introduction to SAS® University Edition - SAS Institute, 2015
7. Geoff Der, Brian S. Everitt Essential Statistics Using SAS® University Edition - SAS Institute, 2015

Ресурсы информационно-телекоммуникационной сети «Интернет»

1. SASUniversityEdition, руководство по установке https://www.sas.com/ru_ru/software/university-edition/download-software.html
2. SAS Programming I: Essentialshttps://support.sas.com/edu/schedules.html?ctry=us&id=277
3. SAS Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression https://support.sas.com/edu/schedules.html?ctry=us&id=1979.
4. Oracle Virtual Box http://www.virtialbox.org

Информационные технологии, используемые в процессе обучения

1. Бесплатная аналитическая платформа SASUniversityEdition
2. Один из следующих браузеров: MicrosoftInternetExplorer версии 9, 10 или 11; MozillaFirefox версии 21 или выше; GoogleChrome версии 27 или выше
3. Бесплатное программное обеспечение для виртуализации OracleVirtualBox

Активные и интерактивные формы проведения занятия

Каждое занятие (лекция и семинар) сопровождается демонстрацией преподавателями изучаемых на данном занятии технологий. В рамках данных демонстраций аспиранты проделывают необходимые действия по настройке программного обеспечения, написанию и запуску программ на своих компьютерах, задают вопросы. Дополнительно, на семинаре аспиранты выполняют небольшие практические задания (как индивидуально, так и в группах) по тематике последней на момент данного семинара лекции. Также на каждом занятии проводится обсуждение домашних заданий, а также все аспиранты имеют возможность задать преподавателям свои вопросы по изучаемой теме.

Материально-техническая база

1. Для преподавания дисциплины требуется класс, оборудованный маркерной или меловой доской и проектором (и компьютером с разъемом VGA / HDMI для подключения к проектору);

2. Для демонстрации современных технологий аналитики больших данных требуется компьютерный класс с доступом в Интернет со следующим установленным набором ПО:
   - Бесплатная аналитическая платформа SAS UniversityEdition
   - Один из следующих браузеров: MicrosoftInternetExplorer версии 9, 10 или 11; MozillaFirefox версии 21 или выше; GoogleChrome версии 27 или выше
   - Бесплатное программное обеспечение для виртуализации OracleVirtualBox.

12. Язык преподавания

Русский.

13. Разработчики программы, Преподаватели

Доцент кафедры Интеллектуальных Информационных Технологий, Петровский Михаил Игоревич (michael@cs.msu.su)

Приложение

Оценочные средства для промежуточной аттестации по дисциплине «ТЕХНОЛОГИИ ПРИКЛАДНОГО АНАЛИЗА ДАННЫХ SAS»

Средства для оценивания планируемых результатов обучения, критерии и показатели оценивания приведены ниже.

РЕЗУЛЬТАТ ОБУЧЕНИЯ по дисциплине (модулю)	**КРИТЕРИИ и ПОКАЗАТЕЛИ ОЦЕНИВАНИЯ РЕЗУЛЬТАТА ОБУЧЕНИЯ по дисциплине (модулю)**					**ОЦЕНОЧНЫЕ СРЕДСТВА**
	1	2	3	4	5	
ЗНАТЬ: современные математические методы, применяющиеся для решения задач в области естественных наук, экономики, социологии и информационно-коммуникационных технологий **Код З1 (ОПК-1)**	Отсутствие знаний	Фрагментарные представления о современных математических методах, применяющихся для решения задач в области естественных наук, экономики, социологии и информационно-коммуникационных технологий	В целом сформированные, но неполные знания о современных математических методах, применяющихся для решения задач в области естественных наук, экономики, социологии и информационно-коммуникационных технологий	Сформированные, но содержащие отдельные пробелы знания о современных математических методах, применяющихся для решения задач в области естественных наук, экономики, социологии и информационно-коммуникационных технологий	Сформированные систематические знания о современных математических методах, применяющихся для решения задач в области естественных наук, экономики, социологии и информационно-коммуникационных технологий	Устный экзамен
УМЕТЬ: применять современные методы постановки и анализа задач в области математики и информатики	Отсутствие умений	Фрагментарные умения применять современные методы	В целом успешное, но не систематическое умение применять современные	Успешное, но содержащее отдельные пробелы умение применять	Сформированное умение применять современные методы постановки и	отчет

Код У1 (ОПК-1)		постановки и анализа задач в области математики и информатики	методы постановки и анализа задач в области математики и информатики	современные методы постановки и анализа задач в области математики и информатики	анализа задач в области математики и информатики	
ВЛАДЕТЬ: навыками оптимального выбора современных методов и средств постановки и анализа задач в области математики и информатики **Код В1 (ОПК-1)**	Отсутствие навыков	Фрагментарное владение навыками оптимального выбора современных методов и средств постановки и анализа задач в области математики и информатики	В целом успешное, но не полное владение навыками оптимального выбора современных методов и средств постановки и анализа задач в области математики и информатики	Успешное, но содержащее отдельные пробелывладение навыками оптимального выбора современных методов и средств постановки и анализа задач в области математики и информатики	Сформированное владение навыками оптимального выбора современных методов и средств постановки и анализа задач в области математики и информатики	отчет
ЗНАТЬ: современные методы построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современныеметоды разработки и реализации алгоритмов их решения **Код З1 (ПК-1)**	Отсутствие знаний	Фрагментарные представления о современных методах построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных	В целом сформированные, но неполные знания о современных методах построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также	Сформированные, но содержащие отдельные пробелы знания о современных методах построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также	Сформированные систематические знания о современных методах построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных	Устный экзамен

		методах разработки и реализации алгоритмов их решения	современных методах разработки и реализации алгоритмов их решения	современных методах разработки и реализации алгоритмов их решения	методах разработки и реализации алгоритмов их решения	
УМЕТЬ: применять современные методы построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современныеметоды разработки и реализации алгоритмов их решения **Код У1 (ПК-1)**	Отсутствие умений	Фрагментарные умения применять современные методы построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современныеметоды разработки и реализации алгоритмов их решения	В целом успешное, но не систематическое умение применять современные методы построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современныеметоды разработки и реализации алгоритмов их решения	Успешное, но содержащее отдельные пробелы умение применять современные методы построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современныеметоды разработки и реализации алгоритмов их решения	Сформированное умение применять современные методы построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современныеметоды разработки и реализации алгоритмов их решения	отчет
ВЛАДЕТЬ: навыками оптимального выбора современных методов построения и анализа математических моделей, возникающих при решении	Отсутствие навыков	Фрагментарное владение навыками оптимального выбора современных	В целом успешное, но не полное владение навыками оптимального выбора	Успешное, но содержащее отдельные пробелывладение навыками оптимального	Сформированное владение навыками оптимального выбора современных	отчет

естественнонаучных задач, а также современных методов разработки и реализации алгоритмов их решения **Код В1 (ПК-1)**		методов построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных методов разработки и реализации алгоритмов их решения	современных методов построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных методов разработки и реализации алгоритмов их решения	выбора современных методов построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных методов разработки и реализации алгоритмов их решения	методов построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных методов разработки и реализации алгоритмов их решения	
ЗНАТЬ: современные алгоритмы разработки программного обеспечения вычислительных комплексов; **Код З1 (ПК-3)**	Отсутствие знаний	Фрагментарные представления о современных алгоритмах разработки программного обеспечения вычислительных комплексов	В целом сформированные, но неполные знания о современных алгоритмах компьютерной математики, о математической теории, лежащей в их основе	Сформированные, но содержащие отдельные пробелы знания о современных алгоритмах компьютерной математики, о математической теории, лежащей в их основе	Сформированные систематические знания о современных алгоритмах компьютерной математики, о математической теории, лежащей в их основе	Устный экзамен
УМЕТЬ: применять современные алгоритмы разработки программного обеспечения	Отсутствие умений	Фрагментарные умения применять современные алгоритмы	В целом успешное, но не систематическое умение применять	Успешное, но содержащее отдельные пробелы	Сформированное умение применять современные алгоритмы	отчет

вычислительных комплексов **Код У1 (ПК-3)**		разработки программного обеспечения вычислительных комплексов	современные алгоритмы разработки программного обеспечения вычислительных комплексов	умениеприменять современные алгоритмы разработки программного обеспечения вычислительных комплексов	разработки программного обеспечения вычислительных комплексов	
ВЛАДЕТЬ: базовыми навыками выбора современных алгоритмов разработки программного обеспечения вычислительных комплексов **Код В1 (ПК-3)**	Отсутствие навыков	Фрагментарноевладение базовыми навыками выбора современных алгоритмов разработки программного обеспечения вычислительных комплексов	В целом успешное, но не полное владение базовыми навыками выбора современных алгоритмов разработки программного обеспечения вычислительных комплексов	Успешное, но содержащее отдельные пробелывладение базовыми навыками выбора современных алгоритмов разработки программного обеспечения вычислительных комплексов	Сформированное владениебазовым и навыками выбора современных алгоритмов разработки программного обеспечения вычислительных комплексов	отчет
ЗНАТЬ: современные методы интеллектуального анализа данных; **Код З1 (ПК-5)**	Отсутствие знаний	Фрагментарные представления о современных методах интеллектуального анализа данных	В целом сформированные, но неполные знания о современных методах интеллектуального анализа данных	Сформированные, но содержащие отдельные пробелы знания осовременных методах интеллектуального анализа данных	Сформированные систематические знания осовременных методах интеллектуального анализа данных	Устный экзамен
УМЕТЬ: применять современные методы интеллектуального анализа данных	Отсутствие умений	Фрагментарные уменияприменять современные методы	В целом успешное, но не систематическое умениеприменять	Успешное, но содержащее отдельные пробелы	Сформированное умение применять современные методы	отчет

Код У1 (ПК-5)		интеллектуального анализа данных	современные методы интеллектуального анализа данных	умениеприменять современные методы интеллектуального анализа данных	интеллектуального анализа данных	
ВЛАДЕТЬ: базовыми навыками выбора методов интеллектуального анализа данных **Код В1 (ПК-5)**	Отсутствие навыков	Фрагментарноевладение базовыми навыками выбора методов интеллектуального анализа данных	В целом успешное, но не полное владение базовыми навыками выбора методов интеллектуального анализа данных	Успешное, но содержащее отдельные пробелывладение базовыми навыками выбора методов интеллектуального анализа данных	Сформированное владениебазовыми навыками выбора методов интеллектуального анализа данных	отчет

Фонды оценочных средств, необходимые для оценки результатов обучения

Список вопросов для устной части экзамена

1. Основные принципы работы шага обработки данных.
2. Работа со структурированными наборами данных и массивами.
3. Форматы и типы данных, процедуры преобразования форматов и типов.
4. Работа с внешними сложно структурированными наборами данных.
5. Организация поиска по ключу с помощью индексов, форматов, хэш-объектов.
6. Пользовательские процедуры и функции.
7. Макропеременные и макроподстановки. Взаимодействие с SAS SQL.
8. Работа с подсистемой вывода.
9. Графические возможности и процедуры аналитической платформы.
10. Процедуры проверки гипотез и дисперсионного анализа.
11. Процедуры построения линейных регрессионных моделей. Смешанные линейные регрессионные модели.
12. Проблема мультиколлинеарности, пошаговый отбор переменных, регуляризация, преобразования пространства признаков.

13. Процедуры поиска главных компонент и кластеризации переменных.
14. Процедуры и инструменты для поиска выбросов.
15. Процедуры построения нелинейных регрессий.
16. Анализ таблиц сопряженности, логистическая регрессия.
17. Обобщенные линейные модели, пуассоновская и гамма регрессии.
18. Сравнение и оценка моделей на тестовом наборе данных.

Примеры вопросов для письменной части экзамена

Письменная часть экзамена охватывает материал всего курса и состоит из заданий следующего типа:

1. Тестовые вопросы с выбором одного варианта ответа из списка предложенных;
2. Тестовые вопросы на выбор нескольких верных утверждений из списка предложенных;
3. Расчетные задачи без выбора вариантов ответа;
4. Задачи на написание программы.

1. Программирование

Напишите макрос на SAS, который получет на вход список текстовых файлов (строку с именами файлов, разделенными пробелами), csvфайл в формате «filename, author, importance» (поле filename – имя файла до 256 символов без пробелов и специальных символов, author - имя автора до 20 символов без пробелов и специальных символов, importance – принимает значения 0 или 1, которое обозначает содержит ли файл важную информацию) и словарь в наборе данных dictionary, который содержит две колонки: word (символьная переменная не более 20 символов) и idf (числовая переменная). Где word – уникальные слова из текстовых файлов (считаем, что все слова на англ., к базовым словоформам приводить не надо, т.е. все слова разные, разделители только пробелы), а idf(inversedocumentfrequency) считается как логарифм отношения общего числа файлов, поделить на число файлов, содержащих слово word: idf(word)=log(count(files)/count(files|word)). Макрос должен сформировать набор данных с именем text_datac векторным представлением входных тектсовых файлов (по одной строке на каждый входной файл) в пространстве слов словаря dictionary, где каждому каждому слову Хиз словаря должна соответсвовать переменная tfidfX, равная частоте встречаемости слова в файле, умноженной на idfслова из словаря (для индексирования переменные можно занумеровать):

tfidf(word,file)= count(word|file)/total_words (file))*idf(word).

Переменнуюtotal_words, filename, author, importancетакженаужновключитьврезультирующийнаборtext_data.

2. Дисперсионный анализ

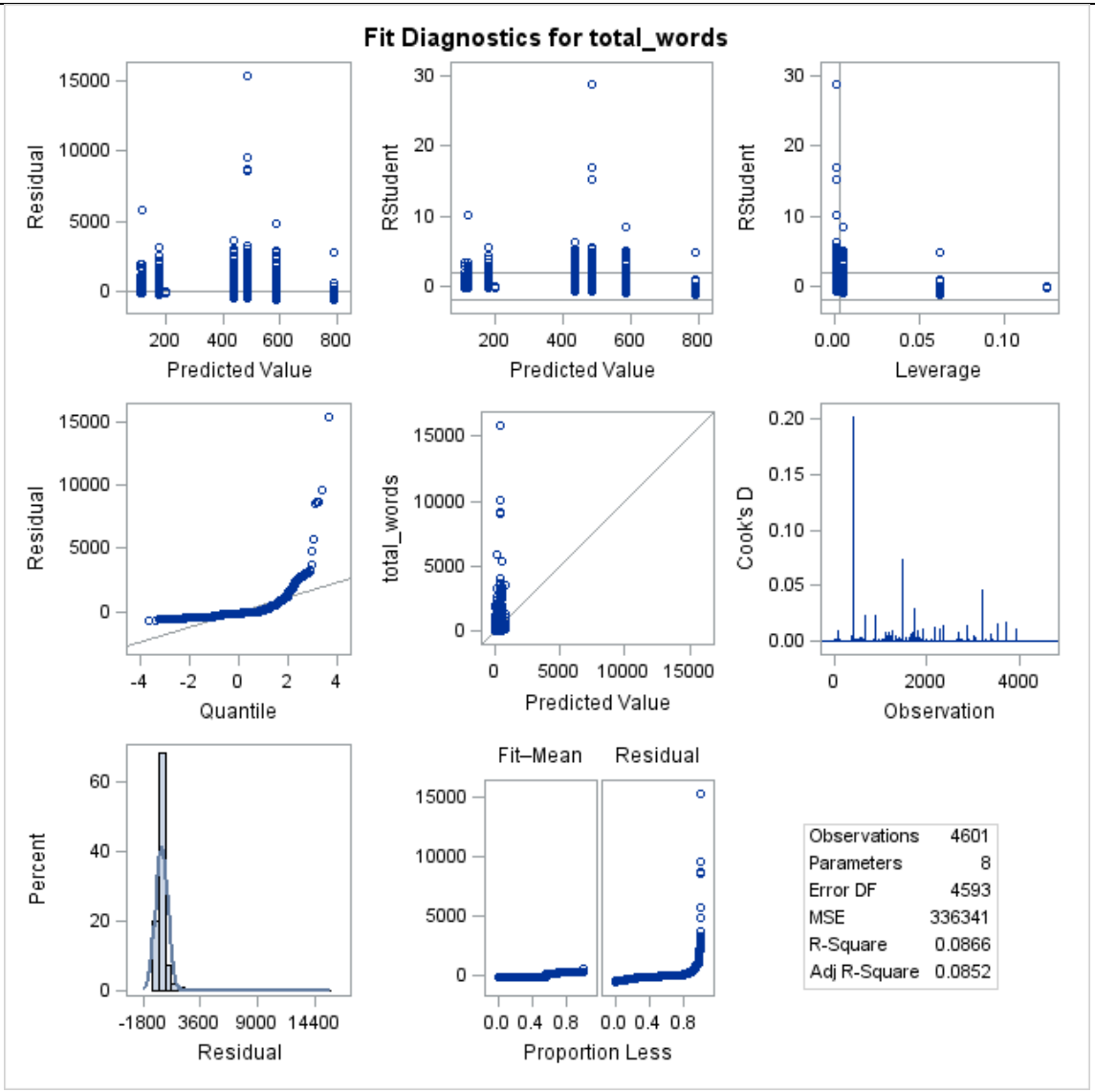
Используя набор данных text_datанаписать программу для проверки предположения, что размер текста в файле (total_words) зависит от авторства и признака, является ли письмо выжным, а также выяснить какие авторы пишут тексты примерно одинакового объема, а какие нет.

Предположим, что Вы получили частичный вывод программы, представленный ниже.

Dependent Variable: total_words

Source	DF	SumofSquares	MeanSquare	F Value	Pr > F
Model	7	146409864	20915695	62.19	<.0001
Error	4593	1544815629	336341		
CorrectedTotal	4600	1691225494			

Source	DF	Type I SS	MeanSquare	F Value	Pr > F
Author	3	29483945.4	9827981.8	29.22	<.0001
importance	1	115367988.8	115367988.8	343.01	<.0001
Author*importance	3	1557930.2	519310.1	1.54	0.2010


Fit Diagnostics for total_words
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
total_words
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 4601
Parameters 8
Error DF 4593
MSE 336341
R-Square 0.0866
Adj R-Square 0.0852

Level of Author	N	total_words	
		Mean	Std Dev
Ivanov	**735**	177.503401	333.115084
Petrov	**203**	601.256158	836.150594
Sidorov	**2439**	296.598196	692.363455
Smirnov	**1224**	267.558007	470.180091

Least Squares Means

Author	total_words LSMEAN	LSMEAN Number
Ivanov	349.046771	1
Petrov	748.973831	2
Sidorov	300.696418	3
Smirnov	271.569527	4

i/j	1	2	3	4
1	_	<.0001	0.0633	0.0067
2	<.0001	_	<.0001	<.0001
3	0.0633	<.0001	_	0.1518
4	0.0067	<.0001	0.1518	_

Ответьтенаследующиевопросы (вездесчитать, чтоуровеньзначимостиравен 0.01):

1) Принята ли базовая гипотеза дисперсионного анализа?

2) Есть ли выбросы в наборе данных

3) Какие предположения дисперсионного анализа нарушены (если нарушены) в данной задаче?

4) Нужно ли использовать в модели переменную Author? Переменную Importance? Их комбинацию?

5) Какие пары авторов неразличимы с точки зрения описания вариации переменной total_words?

3. Регрессионный анализ.

Предположим, что изначально в наборе данных примеров не важных писем было очень много – 99% от выборки. Далее с помощью подхода oversamplingи процедуры surveyselectвыборка была сбалансирована, т.е. получен набор balanced_text_data, где пропорция важных и обычных текстов уже 1:1. На наборе данных balanced_text_data постройте и сохраните модель на основе логистической регрессии для прогнозирования признака, является ли текст важным. При этом должен быть осуществлен отбор значимых переменных комбинированным пошаговым (stepwise) методом. Порог уровня значимости при добавлении переменной должен быть 0.01, а при удалении 0.05. Должны быть выведены ROCкривые для каждого шага. Напишите программу, которая применит полученную модель к набору данных той же структуры с именем score_text_data, где в переменной p_importancebудет записана корректная с учетом балансировки тренировочного набора вероятность того, что текст является важным.

Предположим, что Вы получили частичный вывод программы, представленный ниже.Ответьте на следующте вопросы (везде считать, что уровень значимости равен 0.01):

1) Принята ли базовая гипотеза регрессионного анализа?

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
LikelihoodRatio	3995.6243	27	<.0001
Score	2487.9121	27	<.0001
Wald	808.9079	27	<.0001

2) Какие из перечисленных переменных можно исключить из модели без существенной потери качества? Если их несколько, то можно ли их исключить все?

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard**	**Wald**	**Pr > ChiSq**
Intercept		1	-3.0090	0.1984	229.9632	<.0001
Author	Ivanov	1	-2.6571	0.3837	47.9466	<.0001
Author	Petrov	1	-0.8639	0.3381	6.5296	0.0106
Author	Sidorov	1	1.7507	0.1733	102.0490	<.0001

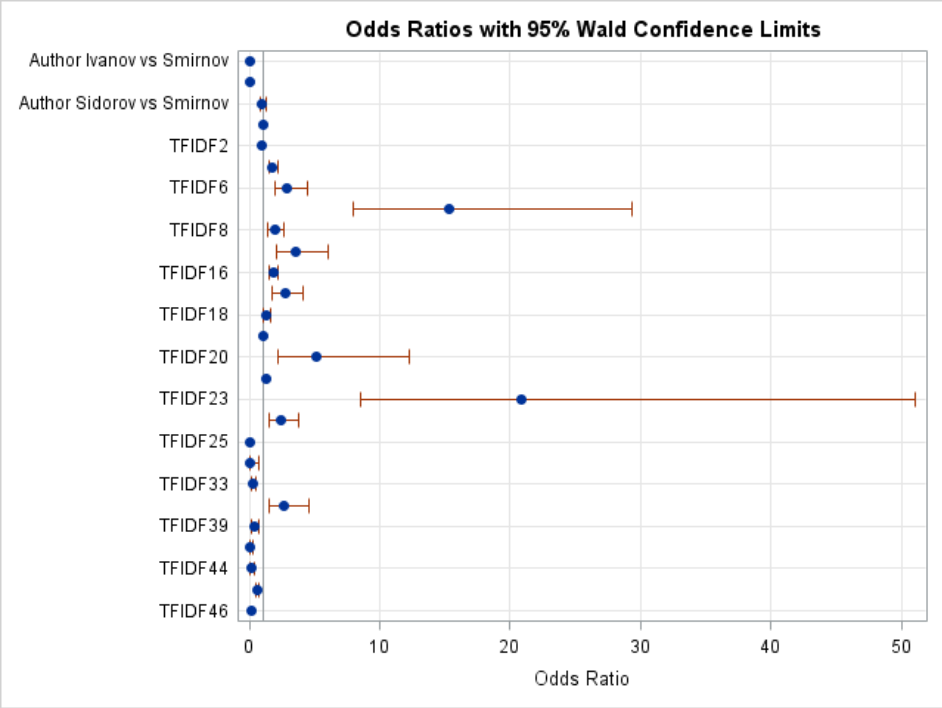
total_words		1	0.00178	0.000187	90.1722	<.0001
TFIDF18		1	0.2710	0.1077	6.3300	0.0119
TFIDF19		1	0.0797	0.0307	6.7381	0.0094
TFIDF20		1	1.6338	0.4437	13.5609	0.0002
TFIDF21		1	0.2166	0.0430	25.3127	<.0001
TFIDF29		1	-3.6589	1.7155	4.5493	0.0329
TFIDF33		1	-1.3133	0.3384	15.0642	0.0001
TFIDF36		1	0.9555	0.2918	10.7218	0.0011

3) Как видно из таблицы ниже процесс отбора переменных остановился на 27 шаге. Почему?

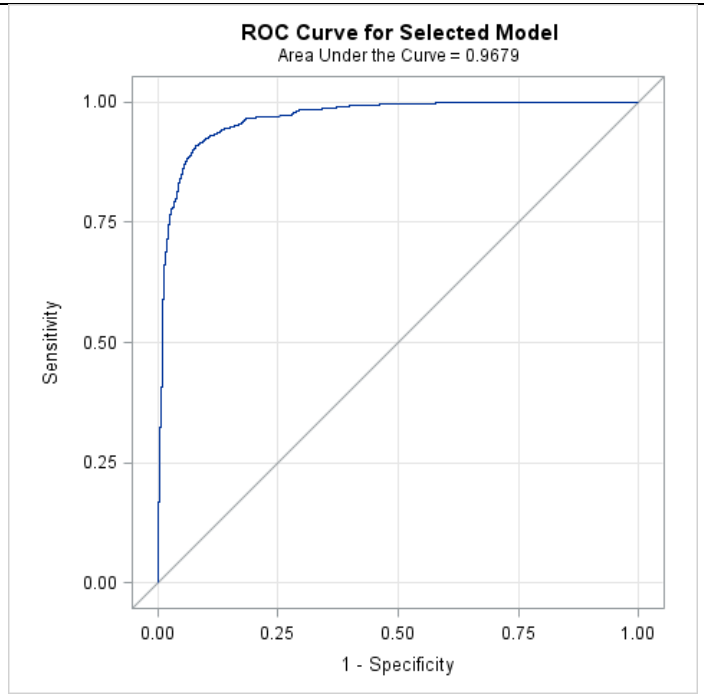
ummaryofStepwiseSelection								
Step	**Effect**		**DF**	**Number In**	**Score Chi-Square**	**Wald Chi-Square**	**Pr > ChiSq**	**Variable Label**
	Entered	**Removed**						
1	**TFIDF21**		1	1	675.7404		<.0001	
2	**Author**		3	2	546.3350		<.0001	
3	**TFIDF23**		1	3	290.8007		<.0001	
4	**TFIDF7**		1	4	289.8012		<.0001	
5	**TFIDF16**		1	5	201.8526		<.0001	
	…	**…**	…	…	…	…	…	…
25	**TFIDF18**		1	25	6.7127		0.0096	
26	**TFIDF47**		1	26	6.6308		0.0100	
27		**TFIDF47**	1	25		3.5285	0.0603	

4) Какая из переменных оказывает наибольшее влияние на отклик? При всех остальных равных переменных, если автором текста является Иванов, то для его

текста вероятность высокой важности ниже чем у Сидорова или выше?



5) Примерно каким будет уровень ложно положительных срабатываний если выбрать порог таким, чтобы не пропустить ни одного важного сообщения? При каком значении ошибки перового и второго рода будут совпадать?



Примеры ПСР приведены выше.

Методические материалы для проведения процедур оценивания результатов обучения

В течение семестра аспиранты выполняют небольшие практические задания на семинарах (по тематике последней на момент данного семинара лекции), а также пять ПСР дома (которые обсуждаются с преподавателями на семинарах и «защищаются»).
За работу на семинарах аспиранты могут получить 0–50 баллов.
За каждую ПСР аспиранты могут получить 0–10 баллов (таким образом, всего за ПСР можно получить 0–40 баллов и еще 10 баллов за необязательную задачу номер 3).
Таким образом, за семестр аспиранты могут набрать 0–100 баллов.
По результатам работы в семестре, всем аспирантам ставится предварительная оценка по следующей схеме:

Количество баллов, набранных в семестре	**Предварительная оценка**
Не менее 80 баллов	«ОТЛ»
Не менее 60 баллов и не более 79 баллов	«ХОР»
Не менее 40 баллов и не более 59 баллов	«УДОВЛ»
Не более 39 баллов	«НЕУД»

Далее, на экзамене аспиранты пишут письменную работу, за которую также получают оценку (вся работа оценивается в 100 баллов, оценка за письменную работу ставится аналогично оценке за работу в семестре).
Итоговая оценка за дисциплину вычисляется как среднее арифметическое между оценкой за работу в семестре и оценкой за письменную на экзамене работу. В случае возникновения спорной ситуации, преподаватели устно задают аспиранту любые три вопроса из списка вопросов для устной части экзамена. По результатам ответа аспиранта на вопросы, ставится итоговая оценка.