Приложение к ОПОП «Информационные системы и технологии в инновационной деятельности»

СПбГЭТУ «ЛЭТИ»
ПЕРВЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ

МИНОБРНАУКИ РОССИИ

федеральное государственное автономное образовательное учреждение высшего образования
«Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И.Ульянова (Ленина)»
(СПбГЭТУ «ЛЭТИ»)

РАБОЧАЯ ПРОГРАММА

дисциплины

«АНАЛИЗ ДАННЫХ»

для подготовки бакалавров

по направлению

27.03.05 «Инноватика»

по профилю

«Информационные системы и технологии в инновационной деятельности»

Санкт-Петербург

2022

ЛИСТ СОГЛАСОВАНИЯ

Разработчики:

д.т.н., профессор Денисов А.Р.

Рабочая программа рассмотрена и одобрена на заседании кафедры ИМ 20.04.2022, протокол № 3

Рабочая программа рассмотрена и одобрена учебно-методической комиссией ИНПРОТЕХ, 27.04.2022, протокол № 7

Согласовано в ИС ИОТ

Начальник ОМОЛА Загороднюк О.В.

1 СТРУКТУРА ДИСЦИПЛИНЫ

Обеспечивающий факультет	ИНПРОТЕХ
Обеспечивающая кафедра	ИМ
Общая трудоемкость (ЗЕТ)	3
Курс	4
Семестр	7
Виды занятий	
Лекции (академ. часов)	17
Практические занятия (академ. часов)	34
Иная контактная работа (академ. часов)	1
Все контактные часы (академ. часов)	52
Самостоятельная работа, включая часы на контроль (академ. часов)	56
Всего (академ. часов)	108
Вид промежуточной аттестации	
Дифф. зачет (курс)	4

2 АННОТАЦИЯ ДИСЦИПЛИНЫ

«АНАЛИЗ ДАННЫХ»

Дисциплина посвящена проблемам использования методов машинного обучения и анализа данных для поддержки принятия управленческих и экономических решений в различных областях жизнедеятельности, в т.ч. при разработке планов инновационного развития предприятий и территорий. Предметом ее изучения являются методические и методологические основы науки об анализе данных. Формирование систематизированного представления о концепциях, моделях и принципах технологий машинного обучения и анализа данных.

SUBJECT SUMMARY

«DATA ANALYSIS»

The discipline is devoted to the problems of using machine learning methods and data analysis to support the adoption of managerial and economic decisions in various areas of life, incl. when developing plans for innovative development of enterprises and territories. The subject of her study is the methodological and methodological foundations of the science of data analysis.

3 ОБЩИЕ ПОЛОЖЕНИЯ

3.1 Цели и задачи дисциплины

1. Цель дисциплины -изучение принципов использования технологий машинного обучения и анализа данных и формирование навыков их применения при принятии управленческих и экономических решений в различных областях жизнедеятельности, в т.ч. при разработке планов инновационного развития предприятий и территорий.

2. Задачи дисциплины:

-изучение технологий машинного обучения и анализа данных, получение представления о консолидации, трансформации данных и способах их визуализации;

– знакомство с типовыми алгоритмами решения профессиональных задач на основе анализа данных;

– развитие умений применять инструментальные средства машинного обучения и анализа данных;

– развитие компетенций в области построения и оценки качества моделей машинного обучения и анализа данных.

3. Формируемые знания:

Предметная область анализа данных

Классификация методов анализа данных

Структура цикла машинного обучения CRISP-DM

Алгоритмы машинного обучения при решении регрессионных, классификацилнных и кластеризационных задач

Методы оценки качества моделей

Современные методы и инструментальные средства анализа данных

Методы интерпретации и визуализации анализа данных.

4. Формируемые умения:

Разрабатывать и оценивать модели данных

Решать задачи кластеризации, регрессии, прогнозирования, снижения размерности и ранжирования данных

Планировать и проводить аналитические работы

Использовать инструментальные средства для извлечения, преобразования, хранения и обработки данных из разнородных источников

Планировать и проводить аналитические работы при решении управленческих и экономических задач в различных областях жизнедеятельности, в т.ч. при разработке планов инновационного развития предприятий и территорий.

5. Формируемые навыки:

Извлечение, проверка, очистка и агрегация данных и разработка представления данных

Оценка соответствия набора данных предметной области и задачам аналитических работ

Разработка, поверка, оценка используемых моделей данных.

3.2 Место дисциплины в структуре ОПОП

Дисциплина изучается на основе ранее освоенных дисциплин учебного плана:

1. «Всеобщий менеджмент качества»

2. «Информатика»

3. «Информационные технологии в логистике»

4. «Информационные технологии в управлении предприятием»

5. «Математический анализ»

6. «Общая теория статистики»

7. «Основы менеджмента качества и управления бизнес-процессами»

8. «Теория вероятностей и математическая статистика»

9. «Экономико-математические методы и модели»

и обеспечивает изучение последующих дисциплин:

1. «Технологии цифровых двойников»

2. «Производственная практика (преддипломная практика)»

3.3 Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы

В результате освоения образовательной программы обучающийся должен достичь следующие результаты обучения по дисциплине:

Код компетенции/ индикатора компетенции	**Наименование компетенции/индикатора компетенции**
УК-10	Способен принимать обоснованные экономические решения в различных областях жизнедеятельности
УК-10.2	*Проводит экономическую оценку и обоснование принимаемых управленческих решений*
ПК-1	Способен участвовать в разработке планов инновационного развития предприятий и территорий
ПК-1.1	*Анализирует проект (инновацию) как объект управления*

4 СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

4.1 Содержание разделов дисциплины

4.1.1 Наименование тем и часы на все виды нагрузки

№ п/п	Наименование темы дисциплины	Лек, ач	Пр, ач	ИКР, ач	СР, ач
1	Введение в анализ данных. Жизненный цикл CRISP-DM	1	4		6
2	Классификация методов анализа данных	2	4		6
3	Работа с распределениями случайных величин	2	4		6
4	Задача регрессии	2	8	1	10
5	Задача классификации	2	4		8
6	Нормализация данных	2	4		6
7	Задача кластеризации. Многомерное сжатие данных	2	4		8
8	Задача авторегрессии	2	2		4
9	DEA-анализ	2			2
	Итого, ач	17	34	1	56
	Из них ач на контроль	0	0	0	0
	Общая трудоемкость освоения, ач/зе	108/3			

4.1.2 Содержание

№ п/п	Наименование темы дисциплины	Содержание
1	Введение в анализ данных. Жизненный цикл CRISP-DM	Назначение систем анализа данных. Причины, обусловившие актуальность данной темы: перманентный реинжиниринг, задача автоматизации интеллектуальных операций, HR, цифровая экономика, понятие BigData: 3V, 5V, 7V. Трехуровневая архитектура системы анализа данных. Сходство и различие в понятиях: Статистика, Эконометрика, Машинное обучение. Цикл машинного обучения: от постановки задачи до принятия решения. Проблема ошибок первого и второго рода. HADI и CRISP-DM. Структура проекта анализа данных: роли в команде. Стадии формирования моделей.
2	Классификация методов анализа данных	Задачи анализа и прогнозирования. Линейные и нелинейные методы. Основные задачи анализа данных: регрессия, классификация, кластеризация. Дополнительные методы: анализ распределений и поиск аномалий, многомерное сжатие, DEA-анализ, распознавание образов, рекомендательные системы и заполнение пропусков, ассоциативные правила. Ансамбли моделей: бэггинг, стекинг, бустинг.

№ п/п	Наименование темы дисциплины	Содержание
3	Работа с распределениями случайных величин	Понятие случайной величины. Законы распределения случайных величин. Базовые законы распределения: распределение Бернулли, нормальный закон и закон Пуассона. Нормальный закон распределения, методы проверки нормальности: критерий Пирсона, критерий Шапиро-Уилка, qqplot. Вероятностный гипотико-дедуктивный подход к решению задач. Алгоритм формулирования и тестирования гипотез. Проблема множественности гипотез. Методы работы с множеством гипотез: методы Холма, Бонферрони, Шидака, Бенджамини. Задача выявления и анализа аномалий.
4	Задача регрессии	Общая постановка задачи регрессии. Задача линейной регрессии. Проблема корреляции входных параметров, регуляризация. Нелинейные методы: понятие дерева решений, случайный лес и градиентный бустинг, К ближайших соседей. Оценка качества регрессии. Требование статичности ошибки: гомосекдастичность и гетероскедастичность, критерии оценки статичности ошибки. Оценка значимости регрессии: R2 и критерий Фишера. Оценка параметров линейной регрессии: критерий стьюдента. Выбор лучшей модели, критерий Акаике.
5	Задача классификации	Общая постановка задачи классификации. Линейные методы классификации: линейная и логистическая регрессия, метод опорных векторов. Использование метода опорных векторов при решении нелинейных задач. Нелинейные методы классификации: случайный лес и градиентный бустинг, К ближайших соседей. Критерии качества результатов классификации: accuracy, precision, recall, f1 метрики. ROC-кривая. Проблема балансировки данных при решении задач классификации. Методы балансировки.
6	Нормализация данных	Принцип GIGO и проблема качества данных. Причины низкого качества данных и методы их выявления. Задача нормализации. Нормализация количественных параметров, нормализация категориальных параметров. Устранение пропусков в данных.
7	Задача кластеризации. Многомерное сжатие данных	Общая постановка задачи кластеризации. Линейные методы кластеризации: k-средних, EM, MeanShift. Выбор лучшей модели по критерию Акаике. Нелинейные методы кластеризации: HDBScan. Анализ результатов кластеризации: визуализация результатов, анализ взаимного расположения множеств, использование логистической регрессии. Задача многомерного сжатия. Линейные методы многомерного сжатия: методы главных компонент и SVD-преобразований. Нелинейные методы многомерного сжатия: MDS и tSNE. Выделение и прогнозирование трендов в компонентах данных.

№ п/п	Наименование темы дисциплины	Содержание
8	Задача авторегрессии	Понятие временного ряда. Задача прогнозирования временного ряда. Выделение компонент временных рядов: трендовая, сезонная и случайная компоненты. Базовые методы прогнозирования временных рядов: авторегрессия и скользящее среднее. Современные методы прогнозирования временных рядов: SARIMA и GARCH. Оценка качества авторегрессионных моделей.
9	DEA-анализ	Понятие эффективности и методы ее оценки. Проблема оценки эффективности многокритериальных задач. Метод анализа среды функционирования (Data envelopment analysis). CRS и VRS модели. Модели ориентированные на достижение максимума эффектов (output-oriented) и минимума ресурсов (input-oriented). Метод выпуклых оболочек (FDH). Оценка устойчивости результатов DEA-анализа.

4.2 Перечень лабораторных работ

Лабораторные работы не предусмотрены.

4.3 Перечень практических занятий

Наименование практических занятий	Количество ауд. часов
1. Основы python. List и алгоритмические структуры	2
2. Основы python. Numpy.array и построение графиков	2
3. Моделирование центральной предельной теоремы	4
4. Формулирование гипотез машинного обучения	2
5. Прогнозирование пола и возраста по фотографии	2
6. Однофакторный регрессионный анализ, линейная регрессия и регуляризаторы	4
7. Авторегрессионная задача прогнозирования финансовых трендов	2
8. Классификационная задача кредитного скоринга	6
9. Кластеризация данных о студентах	4
10. Задача прогнозирования аренды велосипедов	6
Итого	34

4.4 Курсовое проектирование

Курсовая работа (проект) не предусмотрены.

4.5 Реферат

Реферат не предусмотрен.

4.6 Индивидуальное домашнее задание

В рамках курса студенты должны построить 3 модели для решения задач анализа данных (регрессионную, классификационную, кластеризационную). Каждая модель оформляется в виде отдельного индивидуального задания:

ИДЗ № 1 (задача регрессии) выдается на 4-ой неделе;

ИДЗ № 2 (задача классификации) выдается на 8-ой неделе;

ИДЗ № 3 (задача кластеризации) выдается на 13-ой неделе.

Выполнение ИДЗ предполагает самостоятельное изучение учебно-методических материалов по дисциплине, а также литературных источников и ресурсов Интернет по темам ИДЗ и подготовка доклада с презентацией. Работа выполняется либо индивидуально, либо командно. Команда может состоять от 2 до 4 человек. Роли в команде, в т.ч. роль лидера, распределяются самими участниками. Итоговые модели проектируются в среде разработки ноутбуков (Jupiter, Colaboratory Google ...).

Результаты загружаются в Moodle. Защита работ осуществляется публично на контрольной неделе.

4.7 Доклад

Доклад не предусмотрен.

4.8 Кейс

Кейс не предусмотрен.

4.9 Организация и учебно-методическое обеспечение самостоятельной работы

Изучение дисциплины сопровождается самостоятельной работой студентов с рекомендованными преподавателем литературными источниками и информационными ресурсами сети Интернет.

Планирование времени для изучения дисциплины осуществляется на весь период обучения, предусматривая при этом регулярное повторение пройденного материала. Обучающимся, в рамках внеаудиторной самостоятельной работы, необходимо регулярно дополнять сведениями из литературных источников материал, законспектированный на лекциях. При этом на основе изучения рекомендованной литературы целесообразно составить конспект основных положений, терминов и определений, необходимых для освоения разделов учебной дисциплины.

Особое место уделяется консультированию, как одной из форм обучения и контроля самостоятельной работы. Консультирование предполагает особым образом организованное взаимодействие между преподавателем и студентами, при этом предполагается, что консультант либо знает готовое решение, которое он может предписать консультируемому, либо он владеет способами деятельности, которые указывают путь решения проблемы.

Текущая СРС	**Примерная трудоемкость, ач**
Работа с лекционным материалом, с учебной литературой	8
Опережающая самостоятельная работа (изучение нового материала до его изложения на занятиях)	0
Самостоятельное изучение разделов дисциплины	8
Выполнение домашних заданий, домашних контрольных работ	30
Подготовка к лабораторным работам, к практическим и семинарским занятиям	0
Подготовка к контрольным работам, коллоквиумам	0
Выполнение расчетно-графических работ	0
Выполнение курсового проекта или курсовой работы	0

Текущая СРС	Примерная трудоемкость, ач
Поиск, изучение и презентация информации по заданной проблеме, анализ научных публикаций по заданной теме	0
Работа над междисциплинарным проектом	0
Анализ данных по заданной теме, выполнение расчетов, составление схем и моделей, на основе собранных данных	0
Подготовка к зачету, дифференцированному зачету, экзамену	10
ИТОГО СРС	56

5 Учебно-методическое обеспечение дисциплины

5.1 Перечень основной и дополнительной литературы, необходимой для освоения дисциплины

№ п/п	Название, библиографическое описание	К-во экз. в библ.
Основная литература		
1	Мхитарян, Владимир Сергеевич. Анализ данных [Электронный ресурс] : Учебник для вузов / под ред. Мхитаряна В.С., 2020. -490 с	неогр.
2	Маккинни У. Python и анализ данных [Электронный ресурс] : научное издание / У. Маккинни, 2020. -540 с.	неогр.
Дополнительная литература		
1	Анализ данных [Электронный ресурс]. Ч. 1 : учебное пособие, 2020. -162 с.	неогр.
2	Миркин, Борис Григорьевич. Введение в анализ данных [Электронный ресурс] : Учебник и практикум / Миркин Б. Г., 2020. -174 с	неогр.

5.2 Перечень ресурсов информационно-телекоммуникационной сети «Интернет», используемых при освоении дисциплины

№ п/п	Электронный адрес
1	Колаборатория Гугл https://colab.research.google.com/
2	Anaconda project https://www.anaconda.com/

5.3 Адрес сайта курса

Адрес сайта курса: https://vec.etu.ru/moodle/course/view.php?id=10707

6 Критерии оценивания и оценочные материалы

6.1 Критерии оценивания

Для дисциплины «Анализ данных» формой промежуточной аттестации является зачет с оценкой.

Зачет с оценкой

Оценка	Описание
Неудовлетворительно	Не выполнено и защищено хотя бы одно ИДЗ из трех
Удовлетворительно	Выполнены и защищены все 3 ИДЗ, не выполнены остальные практические работы
Хорошо	Выполнены и защищены все ИДЗ и практические работы, студент не смог ответить на теоретический вопрос
Отлично	Выполнены и защищены все ИДЗ и практические работы, студент ответил на теоретический вопрос

Особенности допуска

Для допуска к зачету студент должен выполнить все ИДЗ. Итоговая оценка ставится на основе результатов защит ИДЗ и практических работ, а также ответа на теоретический вопрос

6.2 Оценочные материалы для проведения текущего контроля и промежуточной аттестации обучающихся по дисциплине

Примерные вопросы к дифф.зачету

№ п/п	Описание
1	Цикл CRISP-DM
2	Задача регрессии. Линейная регрессия. Регуляризаторы
3	Задача регрессии. Деревья решений. Алгоритм случайного леса
4	Задача регрессии. Деревья решений. Алгоритм градиентного бустинга
5	Задача регрессии. Метод К ближайших соседей
6	Задача регрессии. Критерии оценки качества регрессии
7	Задача регрессии. Оценка гетероскедастичности
8	Авторегрессионная модель
9	Анализ распределний. Нормальное распределение
10	Задача классификации. Логистическая регрессия
11	Задача классификация. Метод опорных векторов
12	Задача классификация. Нелинейные модели
13	Задача классификации. Оценка качества модели. Accuracy, precision, recall, f метрики
14	Задача классификации. Оценка качества модели. ROC-кривая, ROC-AUC метрика
15	Комбинирование моделей: бустинг, стэкинг, бэггинг
16	Проблема балансировки данных. Методы балансировки
17	Нормализация данных. Нормализация количественных и категориальных параметров. Борьба с пропусками
18	Задача кластеризации. Линейные и нелинейные методы кластеризации
19	Задача кластеризациии. Анализ и интерпретация результатов кластеризации
20	DEA-анализ. CRS и VRS модели
21	DEA-анализ. FDH модели
22	Многомерное сжатие. Модель главных компонент
23	Многомерное сжатие. MDS и tSNE модели

Форма билета

Министерство науки и высшего образования Российской Федерации

ФГАОУ ВО «Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» имени В.И. Ульянова (Ленина)»

БИЛЕТ К ЗАЧЕТУ № 1

Дисциплина ”Анализ данных” ИНПРОТЕХ

1. Цикл CRISP-DM.

УТВЕРЖДАЮ

Заведующий кафедрой И.А. Брусакова

Весь комплект контрольно-измерительных материалов для проверки сформированности компетенции (индикатора компетенции) размещен в закрытой части по адресу, указанному в п. 5.3

6.3 График текущего контроля успеваемости

Неделя	Темы занятий	Вид контроля
4	Задача регрессии	
5		
6		
7		ИДЗ / ИДРГЗ / ИДРЗ
8	Задача классификации	
9		
10		
11		
12		ИДЗ / ИДРГЗ / ИДРЗ
13	Задача кластеризации. Многомерное сжатие данных	
14		
15		
16		
17		ИДЗ / ИДРГЗ / ИДРЗ

6.4 Методика текущего контроля

Методика текущего контроля на лекционных занятиях

Текущий контроль осуществляется через способность студентов применять теоретические концепты, изложенные в лекциях, при выполнении пратических и индивидуальных заданий.

Методика текущего контроля на практических (семинарских) занятиях

Текущий контроль включает в себя:

- подготовку и защиту результатов индивидуальных заданий. Тема заданий выдается преподавателем. Работа выполняется или индивидуально, или командно от 2 до 4 человек, роли в команде распределяют сами участники.

- проверку практических работ.

Критерии оценки индивидуальных заданий

1. Регрессионная модель

- Проведена нормализация количественных и категориальных параметров

- Правильно подобрана модель для тренда

- Проведено исследование и выбрана лучшая модель для учета входных факторов

- Оценена необходимость выявления регрессионной составляющей, в случае необходимости правильно подобрана модель

- Правильно подобраны и использованы критерии оценки качества модели.

2. Классификационная модель

- Проведена нормализация количественных и категориальных параметров

- Правильно подобран метод балансировки днных

- Проведено исследование и выбрана лучшая модель для классификации

- Правильно подобраны и использованы критерии оценки качества модели.

3. Кластеризационная модель

- Проведена нормализация количественных и категориальных параметров

- Проведено исследование и выбрана лучшая модель для кластеризации

- Проведено исследование по интерпретации результатов кластеризации

Оценка за практические работы и ИДЗ выставляется по четырех-балльной шкале по следующим критериям:

«отлично» работа выполнена полностью;

«хорошо» работа выполнена частично;

«удовлетворительно» в работе имеются существенные ошибки;

«неудовлетворительно» работа не выполнена или выполнена неправиль-

но.

Методика текущего контроля самостоятельной работы студентов

Контроль самостоятельной работы студентов осуществляется на практических занятиях студентов по результатам выполнения индивидуальных заданий.

7 Описание информационных технологий и материально-технической базы

Тип занятий	Тип помещения	Требования к помещению	Требования к программному обеспечению
Лекция	Лекционная аудитория	Количество посадочных мест – в соответствии с контингентом, рабочее место преподавателя, проектор, экран, маркерная доска, ПК	1) Windows XP и выше; 2) Microsoft Office 2007 и выше
Практические занятия	Компьютерный класс	Количество посадочных мест – в соответствии с контингентом, рабочее место преподавателя, проектор, экран, маркерная доска, ПК	1) Windows XP и выше; 2) Microsoft Office 2007 и выше
Самостоятельная работа	Помещение для самостоятельной работа	Оснащено компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду университета.	1) Windows XP и выше; 2) Microsoft Office 2007 и выше

8 Адаптация рабочей программы для лиц с ОВЗ

Адаптированная программа разрабатывается при наличии заявления со стороны обучающегося (родителей, законных представителей) и медицинских показаний (рекомендациями психолого-медико-педагогической комиссии). Для инвалидов адаптированная образовательная программа разрабатывается в соответствии с индивидуальной программой реабилитации.

ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ

№ п/п	Дата	Изменение	Дата и номер протокола заседания УМК	Автор	Начальник ОМОЛА